3 WAY EMOTION DETECTION

Project Reference No.: *47S_BE_4321*

College ***:*** *Jyothy Institute of Technology, Bengaluru*
Branch ***:*** *Department of Information Science and Engineering*
Guide(s) ***:*** *Mr. Srinidhi Kulkarni*
Student(S) ***:*** *Mr. Gnanesh CS*
Mr. Punith SY
Mr. Sanjay HP
Ms. Sanjana A Hiremath

INTRODUCTION

In the fast-paced and interconnected world of today, understanding and responding to human emotions have become pivotal for enhancing user experience and well-being. This project embarks on the development of an advanced emotion detector that harnesses the power of three key modalities: facial expressions, voice characteristics, and textual content. Recognizing that emotions manifest in multifaceted ways, we employ Convolutional Neural Networks (CNNs) to analyze facial expressions, librosa and CNNs for voice emotion detection using Mel-Frequency Cepstral Coefficients (MFCC), and natural language processing (NLP) coupled with machine learning for extracting sentiments from textual data. By synergistically combining these technologies, we aim to create a more holistic and accurate representation of the user's emotional state.

The primary goal of this project extends beyond mere emotion detection; it endeavors to leverage the detected emotional states to provide personalized recommendations for the user. These recommendations are designed to cater to the user's emotional needs, offering suggestions such as nearby healthcare professionals or therapists based on detected emotions, curated movie and podcast selections aligned with the user's current emotional state, and tailored news articles to maintain a balanced and informed perspective. In addition, the project integrates with YouTube, offering emotion-aware video recommendations, thereby creating a unique and empathetic user experience that contributes positively to their mental and emotional health. Through this project, we aspire to showcase the potential of integrating diverse technologies to create a sophisticated, emotion-aware system that enhances user engagement and satisfaction in a variety of domains.

1.1 Motivation

In an increasingly interconnected world, the importance of mental health and emotional well- being cannot be overstated. Despite advancements in technology and

heightened awareness of mental health issues, many individuals still struggle to effectively manage their emotions and seek appropriate support. Traditional methods of emotion detection and support often lack the depth and personalization required to address the complex emotional needs of individuals.

Recognizing this gap, there is a pressing need for more sophisticated and user- centric approaches to emotional well-being.

The motivation behind this project lies in addressing the shortcomings of existing emotion detection systems and providing a comprehensive solution that integrates cutting-edge technologies. By harnessing the power of facial expression analysis, voice emotion detection, and textual sentiment analysis, we aim to develop a holistic understanding of the user's emotional state. This multifaceted approach not only captures a wider range of emotional cues but also enables more accurate and personalized support for individuals navigating their emotional landscape.

Furthermore, this project is motivated by the potential of technology to bridge the gap between individuals and the support they need. By leveraging machine learning algorithms and natural language processing techniques, we aspire to create a supportive ecosystem that empowers individuals to better understand and manage their emotions. Through personalized recommendations for emotional support resources, curated media content, and empathetic engagement, we aim to foster emotional resilience and flourishing in individuals across diverse communities.

Moreover, this project aligns with the broader societal shift towards prioritizing mental health and well-being. As conversations around mental health become increasingly normalized, there is a growing demand for innovative solutions that cater to individual needs and preferences. By developing an emotion detection system that combines advanced technology with empathetic design principles, we seek to contribute to the advancement of user-centric technology solutions that prioritize emotional well-being.

Ultimately, the motivation behind this project stems from a shared vision of creating a more inclusive and supportive digital landscape that fosters emotional well-being for all. By leveraging the power of technology to gain deeper insights into human emotions and provide personalized support, we aspire to make a meaningful impact on the lives of individuals, empowering them to lead healthier and more fulfilling lives.

1.2 Existing System

Current emotion detection systems often rely on isolated modalities, such as facial expression analysis or voice emotion detection, resulting in a limited understanding of the complex nature of human emotions. These systems typically lack integration across modalities and fail to provide personalized recommendations based on the detected emotions. Additionally, many emotion detectors lack the sophistication to interpret emotions from textual content, overlooking a crucial dimension of user emotion. The absence of a comprehensive and integrated approach in the existing systems hinders their ability to deliver a nuanced understanding of user emotions and, consequently, limits their effectiveness in providing tailored recommendations for user engagement and well-being.

DRAWBACKS

1. Limited scope due to unimodal focus, neglecting the holistic nature of human emotions.
2. Lack of integration across modalities hampers accuracy in emotion detection.
3. Absence of personalized recommendations based on detected emotions limits user engagement and well-being enhancement.

1.3 Proposed System

The proposed system aims to overcome the limitations of existing emotion detection frameworks by integrating facial expression analysis, voice emotion detection, and text sentiment analysis into a unified and sophisticated model. Leveraging Convolutional Neural Networks (CNNs) for facial expressions, librosa for voice features, and Natural Language Processing (NLP) for textual content, the system ensures a comprehensive understanding of user emotions. By seamlessly combining these modalities, our approach provides a holistic and nuanced representation of emotional states. Moreover, the proposed system goes beyond detection, offering personalized recommendations based on the user's emotions, including suggestions for nearby healthcare professionals, tailored movie and podcast selections, curated news articles, and emotion-aware YouTube content. This holistic and user- centric approach is designed to significantly enhance the overall user experience, fostering emotional well-being through empathetic and contextually relevant content recommendations.

ADVANTAGES

1. Comprehensive emotion understanding through integration of facial expressions, voice features, and textual sentiment analysis.
2. Holistic recommendations enhance user experience, offering personalized content alignedwith detected emotions.
3. Improved accuracy and relevance in emotion detection and content recommendations dueto the integrated multimodal approach.

1.4 Objectives

1. Develop a robust Convolutional Neural Network (CNN) model for facial expression analysis to accurately detect and classify a range of emotions from images.
2. Implement voice emotion detection using the librosa library, integrating CNNs and Mel- Frequency Cepstral Coefficients (MFCC) for precise emotion extraction from audio signals.
3. Employ Natural Language Processing (NLP) techniques and machine learning algorithms to analyze and classify emotional sentiment in textual content for comprehensive emotion understanding.
4. Integrate the three modalities into a cohesive system, enabling real-time emotion detection and providing personalized recommendations for healthcare professionals, movies, podcasts, news, and YouTube content based on the user's emotional state.

1.5 Features and Scope

The system encompasses a comprehensive approach to emotion detection, integrating three primary modalities: facial expressions, voice characteristics, and textual analysis. Each modality offers unique features and capabilities to capture emotional cues from different sources. The scope of facial expression analysis involves utilizing Convolutional Neural Networks (CNNs) to analyze images and extract nuanced emotional signals. This feature enables the system to recognize a wide range of emotions, providing real-time feedback on the user’s emotional state based on visual cues.

Voice emotion detection is a key component of the system, employing CNNs and Mel- Frequency Cepstral Coefficients (MFCC) for audio processing.

The scope of this feature includes analyzing voice characteristics such as tone, pitch, and speech patterns to discern emotional nuances conveyed through speech. By extracting emotional cues from audio recordings, the system enhances its ability to detect and interpret the user's emotional state across diverse communication channels.

Textual sentiment analysis further expands the system's capabilities by utilizing Natural Language Processing (NLP) techniques to analyze textual content for emotional sentiment and mood. The scope of this feature encompasses processing text from various sources, including social media posts, messages, and articles, to gain insights into the user's emotional expressions through written communication. By analyzing textual data, the system gains a deeper understanding of the user's emotional state, complementing the information obtained from facial expressions and voice characteristics.

Integration and fusion of modalities play a crucial role in the system's functionality, enabling the seamless combination of data from multiple sources to gain a holistic understanding of the user's emotional state. Machine learning algorithms are leveraged to integrate data from facial expressions, voice characteristics, and textual content, allowing the system to infer the user's overall emotional state accurately. The scope of this feature includes developing algorithms that can fuse data from different modalities to provide personalized recommendations and interventions tailored to the user's unique emotional needs.

Personalized recommendations constitute a significant aspect of the system, offering tailored interventions to enhance user well-being based on detected emotions. The scope of this feature involves suggesting nearby healthcare professionals or therapists for emotional support, curating media content aligned with the user's mood, and providing tailored news articles. Integration with platforms such as YouTube further enriches user engagement by offering emotion-aware video recommendations, creating a user-centric platform that prioritizes emotional and mental well-being.

1.6 Limitations:

1. Dependency on External Factors: The accuracy of emotion detection can be influenced by external factors such as lighting conditions for facial expression analysis, background noise for voice emotion detection, and the context of textual content for sentiment analysis.
2. Privacy and Data Security Concerns: The collection and processing of personal data,

including facial images, voice recordings, and textual content, raise concerns about privacy and data security. Ensuring the confidentiality and ethical handling of user data is essential but challenging.

3. Emotion Ambiguity and Subjectivity: Emotions are inherently subjective and can vary significantly between individuals. The system may struggle to accurately interpret emotions in cases where expressions are ambiguous or subject to individual interpretation.
4. Limited Contextual Understanding: While the system analyses facial expressions, voice characteristics, and textual content, it may lack contextual understanding of the user's emotional state. Understanding situational context and personal nuances is crucial for providing truly personalized recommendations and interventions.
5. Scalability and Generalization: Adapting the system to different cultural contexts, languages, and user demographics poses challenges in terms of scalability and generalization. Customizing the system to cater to diverse user needs while maintaining accuracy and reliability across various contexts is a complex task.

1.7 Organization of Report

The organization of a report is crucial because it determines how well the content is presented, and how easy it is for the readers to understand and follow the flow of information. A well- organized report provides clarity, structure, and coherence, making it easier for the readers to comprehend the research objectives, methodology, findings and conclusions. The report is organized as follows:

- **Chapter-1: Introduction:** In this chapter, the report provides an overview of the project and its organization. It introduces the topic of Comprehensive Emotion Detector for Personalized well-being Enhancement.

  It provides an overview of the significance of using CNN, Librosa, LSTM models to efficiently explore the rapid data generation in the current market.

- **Chapter-2: Literature Survey:** This chapter presents a comprehensive review of the existing system, including its general working features and their types (if any) with description. It also includes the limitations of existing systems.

- **Chapter-3: System Requirement Specification:** This chapter includes an analysis and feasibility study of the project. It outlines the hardware and software

requirements for the system along with their required versions.

- **Chapter-4: System Design:** This chapter provides an architectural representation of the proposed system. It includesstate diagrams, sequence diagrams, and flow charts. Each diagram is properly titled and explained.

- **Chapter-5: Implementation and Testing:** This chapter discusses the general implementation details of the project and the test setupused or testing. It also includes test cases presented in a tabulated format. The report explains the algorithms used in the implementation.

- **Chapter-6: Results and Discussion:** This chapter presents the results of the project, including screens with discussions. It may include graphs, tables, and other visual aids to illustrate the findings and facilitate discussions on the outcomes.

- **Chapter-7: Conclusion:** In this chapter, the report provides a concise conclusion based on the objectives of the project. It may include 1 or 2 paragraphs summarizing the findings and outcomes of the project.

LITERATURE SURVEY

Although there are many researches works on online voting systems, here we have critically analyzed and summarized twenty research works and projects which are more relevant, recent and pertinent. It is observed that most the recent works address the issue of online voting and use of various information technologies.

2.1 General Working features of the Existing System

The existing system for emotion detection in music utilizes a combination of deep learning techniques, specifically convolutional neural networks (CNNs), and transfer learning models to associate human emotions with music playback. Two primary CNN models are employed: a five-layer model and a global average pooling (GAP) model. These CNNs are designed to analyze facial cues and visual information to detect emotions accurately. Additionally, the system incorporates transfer learning with three pre-trained models: ResNet50, SeNet50, and VGG16. Transfer learning allows the system to leverage knowledge from pre-trained models and adapt it to the specific task of emotion detection in music.

The system's architecture enables it to efficiently process and interpret emotional cues from users' facial expressions and visual inputs. By combining multiple CNN models, it can capture a wide range of emotional states with high accuracy. Transfer learning further enhances performance by leveraging the features learned from large datasets, reducing the need for extensive training on emotion-specific data.

One notable feature of the existing system is its ability to correlate detected emotions with suitable music selections. Music has a profound impact on human emotions, and by associating specific emotional states with corresponding songs, the system enhances user experiences. This integration of emotion detection with music playback provides a personalized and immersive user experience tailored to individual emotional preferences.

Moreover, the system's performance is comparable to state-of-the-art models in emotion detection while being more efficient in terms of computational resources and processing time.

The combination of CNNs and transfer learning allows for robust emotion detection without sacrificing efficiency. This balance between accuracy and efficiency is crucial for real- world applications, where responsiveness and resource utilization are essential considerations.

Overall, the existing system represents a significant advancement in emotion detection technology, particularly in its application to music playback. By leveraging deep learning and transfer learning techniques, it offers accurate and efficient detection of human emotions, enabling personalized and emotionally resonant user experiences in various contexts.

2.2 Different Types

This project may involve the integration of these individual models into a unified system architecture for multi-modal emotion detection. This integration could include techniques such as ensemble learning, where predictions from multiple models are combined to make final decisions, or fusion approaches that combine features extracted from different modalities for improved performance.

Facial Expression Analysis:

Convolutional Neural Networks (CNNs): CNNs are utilized to analyse facial expressions and recognize emotions from images. This involves designing and training deep learning models specifically tailored for facial expression recognition tasks. CNNs are chosen for their ability to effectively capture spatial dependencies in image data, making them well-suited for tasks like facial recognition.

Voice Characteristics Assessment:

CNNs with Mel-Frequency Cepstral Coefficients (MFCC): CNNs are employed along with MFCC concepts for audio processing to extract emotional cues from voice data. MFCC is a widely used feature extraction technique in speech processing due to its effectiveness in capturing the relevant characteristics of audio signals. By combining CNNs with MFCC, the

system can analyze voice characteristics and infer emotional states from audio inputs.

Text Sentiment Analysis:

Natural Language Processing (NLP) Models: Various NLP models and machine learning techniques are used to analyze textual content for emotional sentiment. This includes sentiment

analysis algorithms such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and support vector machines (SVMs). These models are trained on labelled text data to classify the emotional sentiment expressed in textual inputs.

2.3 Related Paper

Table 2.1: Related Papers

SL.NO.	Paper Title	Problem Statement	TechniqueUsed	Outcomes
01.	Human–Robot Collaboration Using Sequential-Recurrent-Convolution-Network-Based Dynamic Face Emotion and Wireless Speech Command Recognitions	Explore human-robot collaboration through dynamic face emotion and wireless speech command recognitions	Introduce a Sequential-Recurrent-Convolution-Network-based approach	Focuses on improving human-robot collaboration using facial and speech-based emotion recognition

SL.NO.	Paper Title	Problem Statement	Technique Used	Outcomes
02.	Cyberbullying Detection Based on Emotion	Enhance cyberbullying detection by integrating emotion and sentiment features, focusing on negative emotions	Juanpablo Heredia, Edmundo Lopes- Silva, and Yudith Cardinale. It was publishedin the year 2022.	Improved detection accuracysurpassing baseline models. Development of robust CDMs for precise identification of harmful online behaviors. Contribution to cyberbullying research, highlighting emotions' role in detection for online
03.	Evaluation of Conversational Agents: Understanding culture, context and environment in emotion detection	Addressing challenges in conversational AI within Black African society, emphasizing the need to explore geographical and cultural factors.	Developing an emotion prediction model using speech and image data with CNN. Using support vector machine models for automatic tagging of prosodic phrases to extract natural language data.	Emotion prediction model achieves accuracies between 85% and 96%. Improvement inF1- score, Precision, and Recall after applying mitigation techniques.

SL.NO.	Paper Title	Problem Statement	Technique Used	Outcomes
04.	Emotion Quantification Using Variational Quantum State Fidelity Estimation	Utilizing variational quantum state fidelity estimation for emotion quantification, Integrating quantum machine learning algorithms for emotion recognition.	Juanpablo Heredia, Edmundo Lopes- Silva, and Yudith Cardinale. It was publishedin the year 2022.	Development of a novel approach for emotion quantification with quantum computing. Advancement in sentiment analysis and emotion recognition methodologies. Exploration of the capabilities and limitations of quantum computing in emotion detection and sentiment analysis.
05.	Adaptive Multimodal Emotion Detection Architecture for Social Robots	Address limitations in existing multimodal emotion recognition for social robots, especially in fusion processes.	Develop a n adaptive emotion recognition architecture for multiple modalities and sources. Embracet Net+	Establish an efficient architecture handling different inputs and data qualities for robots with diverse sensory capacities.

SL.NO.	Paper Title	Problem Statement	TechniqueUsed	Outcomes
06.	Human–Robot Collaboration Using Sequential-Recurrent-Convolution-Network- Based Dynamic Face Emotion and Wireless Speech Command Recognitions	Addressing the challenge of human- robot collaboration through dynamic face emotion and wireless speech command recognition, crucialfor effective interaction.	Developmentof a Sequential-Recurrent-Convolution-Network-based systemintegrating CNN and LSTM for accurate face emotion and speech command recognition.	Achieving real-time recognition capabilities and improved omnidirectional service robot performance, contributing to smoother and more efficient human- robot collaboration.
07.	SCEP—A New Image Dimensional Emotion Recognition Model Based on Spatial and Channel-Wise Attention Mechanisms	The paper tackles the challenge of improving image emotion recognitionby introducing SCEP, a novel model integrating spatial and channel-wise attention mechanisms.	SCEP integrates attention mechanisms for better emotion recognition, emphasizing saliency and object detection. It combines computational modeling for feature extraction with semantic information.	SCEP advances image emotion recognition with improved accuracy through attention mechanisms and computational modeling, offering insightsinto effective feature extraction and predictive modeling.

08.	Context-AwareEmotion Recognition Based on VisualRelationship Detection	The paper aims to improve emotion recognition accuracy by integrating visual relationship detection, considering contextual cues.	The paper integrates visual relationship detection into emotion recognition usingdeep learning for feature extraction	The research improves emotion recognition accuracy by integrating contextual cues from human-object interactions through visual relationship detection.

SL.NO.	Paper Title	Problem Statement	Technique Used	Outcomes
09.	Robust Emotion Recognition Across Diverse Scenes: A Deep NeuralNetwork Approach Integrating Contextual Cues	Addressing the challenge of robust emotion recognition across diverse scenes, emphasizing the integration of contextual cues.	Utilizing a deep neural network approach that integrates contextual cues for emotion recognition, involving feature extraction from faces, objects, and scene context, and employing techniques like feature fusion, object detection,and pose recognition.	Enhancing emotion recognition accuracy across diverse scenes by integrating contextual cues.The deep neuralnetwork modeleffectively incorporates scene context, improving performance in emotion recognition tasks.
10.	Fine-Grained Emotions Influence on Implicit Hate Speech Detection	Emotion recognition lacks robustness across diverse scenes dueto a lack of integration of contextual cues.	Developed a deep neural network approach integrating contextual cues for robust emotion recognition.	Achieved robust emotion recognition across diverse scenes using a deep neural network. Improved performance by integrating contextual cuesand feature fusion. Enhanced emotion recognition accuracy by considering scene context.

SL.NO.	Paper Title	Problem Statement	TechniqueUsed	Outcomes
11.	Emotion Detection From Micro-Blogs Using Novel Input Representation	Investigate methods for emotion detection in micro-blogs	Propose a novel input representation for micro-blog data	Focused on micro-blog data and text-based emotion detection
12.	Utilizing External Knowledge to Enhance Semantics in Emotion Detection in Conversation	Enhance semantics in emotion detection during conversations	Incorporate external knowledge for improved semantic understanding	Concentrates on enhancing emotion detection in conversational contexts through external knowledge integration
13.	Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features	Achieve happy emotion recognition from unconstrained videos	Utilize 3D hybrid deep features for improved video- based emotion analysis	Focus on recognizing happiness in diverse video content
14.	Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model	Conduct sentiment analysis and emotion detection on cryptocurrency-related tweets	Employ an ensemble LSTM- GRU model for enhanced analysis	Investigates emotions and sentiments in the context of cryptocurrency-related social media content

SL.NO.	Paper Title	Problem Statement	TechniqueUsed	Outcomes
15.	Modified Earthworm Optimization With Deep Learning Assisted Emotion Recognition for Human Computer Interface	Enhance emotion recognition for human-computer interface using modified earthworm optimization and deep learning	Combine optimization techniques and deep learning for improved emotion recognition	Addresses emotion recognition in the context of human-computer interface design

[1] "Emotion Detection From Micro-Blogs Using Novel Input Representation", F. Anzum and M. L. Gavrilova, in IEEE.

Emotion is a natural intrinsic state of mind that drives human behavior, social interaction, and decision-making. Due to the rapid expansion in the current era of the Internet, online social media (OSM) platforms have become popular means of expressing opinions and communicating emotions. With the emergence of natural language processing (NLP) techniques powered by artificial intelligence (AI) algorithms, emotion detection (ED) from user-generated OSM data has become a prolific research domain. However, it is challenging to extract meaningful features for identifying discernible patterns from the short, informal, and unstructured texts that are common on micro-blogging platforms like Twitter. In this paper, we introduce a novel representation of features extracted from user-generated Twitter data that can capture users' emotional states. An advanced approach based on Genetic Algorithm (GA) is used to construct the input representation which is composed of stylistic, sentiment, and linguistic features extracted from tweets. A voting ensemble classifier with weights optimized by a GA is introduced to increase the accuracy of emotion detection using the novel feature representation. The proposed classifier is trained and tested on a benchmark Twitter emotion detection dataset where each sample is labeled with either of the six classes: sadness, joy, love, anger, fear, and surprise. The experimental results demonstrate that the proposed approach outperforms the state-of-the art classical machine learning-based emotion detection techniques, achieving the highest level of precision (96.49%), recall (96.49%), F1-score (96.49%), and accuracy (96.49%).

[2] "Robust Emotion Recognition Across Diverse Scenes: A Deep Neural Network Approach Integrating Contextual Cues," X. Zhang, G. Qi, X. Fu and N. Zhang, in IEEE Access.

The emotional context of a given environment can profoundly influence an individual's feelings and responses. Nonetheless, current emotion recognition methodologies primarily concentrate on analyzing the target subject's features and inadequately integrate these features with the contextual information of the scene. To tackle this challenge, we introduce a novel emotion recognition model that employs three independent and prioritized deep convolutional neural networks, alongside a feature fusion enhancement technique, to effectively merge facial information, body pose information, and subject features within the overall image. By amalgamating the performance of object detection models and deep convolutional network models, our framework capitalizes on the strengths of multiple approaches. Experiments with the Emotic dataset validate that our proposed model is technically innovative and surpasses existing methods and benchmark models in terms of feature fusion performance. Moreover, our evaluation of the proposed method on the Emotic dataset underscores the significance of environmental contextual information in shaping human emotions.

[3] "Fine-Grained Emotions Influence on Implicit Hate Speech Detection," A. R. Jafari, G. Li, P. Rajapaksha, R. Farahbakhsh and N. Crespi, in IEEE Access.

Recent years brought an exponential growth of social media which revolutionized freedom of speech but significantly increased the propagation of hate speech and hate-based activities. Therefore, constructive countermeasures are necessary to prevent escalating hateful content on online social media. Many recent works target explicit hate speech, but only a few studies have utilized multiple fused features such as sentiment, targets, and emotions as attributes to enhance the detection of hate speech. In general, sentiment features help to discern feelings such as positivity or negativity, and emotion features provide a deeper level of granularity, focusing on a more comprehensive understanding of sensitivities. The aim of this paper is to investigate the significance of incorporating fine-grained emotions as an essential feature in improving the classification of implicit hate speech. First, we analyzed emotion variations of hateful and non- hateful content and explored their major fine-grained emotion discrepancies targeting implicit hateful content. Next, we introduce a multi-task learning approach that integrates emotions and sentiment features to classify implicit

expressions of hatred. To evaluate the effectiveness of our multi-task learning approach, we compared it with baseline models using single-task learning approaches. The experimental results show that our multi-task approach outperformed in classifying implicit hate speech compared to the baseline models and demonstrates that fine-grained emotional knowledge decreases the classification error across multiple implicit hate categories.

[4] "FF-BTP Model for Novel Sound-Based Community Emotion Detection," A. M. Yildiz et al, in IEEE Access.

Most emotion classification schemes to date have concentrated on individual inputs rather than crowd-level signals. In addressing this gap, we introduce Sound-based Community Emotion Recognition (SCED) as a fresh challenge in the machine learning domain. In this pursuit, we crafted the FF-BTP-based feature engineering model inspired by deep learning principles, specifically designed for discerning crowd sentiments. Our unique dataset was derived from 187 YouTube videos, summing up to 2733 segments each of 3 seconds (sampled at 44.1 KHz). These segments, capturing overlapping speech, ambient sounds, and more, were meticulously categorized into negative, neutral, and positive emotional content. Our architectural design fuses the BTP, a textural feature extractor, and an innovative handcrafted feature selector inspired by Hinton's FF algorithm. This combination identifies the most salient feature vector using calculated mean square error. Further enhancements include the incorporation of a multilevel discrete wavelet transform for spatial and frequency domain feature extraction, and a sophisticated iterative neighborhood component analysis for feature selection, eventually employing a support vector machine for classification. On testing, our FF- BTP model showcased an impressive 97.22% classification accuracy across three categories using the SCED dataset. This handcrafted approach, although inspired by deep learning's feature analysis depth, requires significantly lower computational resources and still delivers outstanding results. It holds promise for future SCED-centric applications.

[5] "Modified Earthworm Optimization With Deep Learning Assisted Emotion Recognition for Human Computer Interface," F. Alrowais et al., in IEEE Access.

Among the most prominent field in the human-computer interface (HCI) is emotion recognition using facial expressions. Posed variations, facial accessories, and non-uniform illuminations are some of the difficulties in the emotion recognition field. Emotion detection with the help of traditional methods has the shortcoming of mutual optimization of feature extraction and classification. Computer vision (CV) technology improves HCI by visualizing

the natural world in a digital platform like the human brain. In CV technique, advances in machine learning and artificial intelligence result in further enhancements and changes, which ensures an improved and more stable visualization. This study develops a new Modified Earthworm Optimization with Deep Learning Assisted Emotion Recognition (MEWODL-ER) for HCI applications. The presented MEWODL-ER technique intends to categorize different kinds of emotions that exist in the HCI applications. To do so, the presented MEWODL-ER technique employs the GoogleNet model to extract feature vectors and the hyperparameter tuning process is performed via the MEWO algorithm. The design of automated hyperparameter adjustment using the MEWO algorithm helps in attaining an improved emotion recognition process. Finally, the quantum autoencoder (QAE) model is implemented for the identification and classification of emotions related to the HCI applications. To exhibit the enhanced recognition results of the MEWODL-ER approach, a wide-ranging simulation analysis is performed. The experimental values indicated that the MEWODL-ER technique accomplishes promising performance over other models with maximum accuracy of 98.91%.

[6] "Cyberbullying Detection Based on Emotion," M. Al-Hashedi, L. -K. Soon, H. -N. Goh, A. H. L. Lim and E. -G. Siew, " in IEEE Access,

Due to the detrimental consequences caused by cyberbullying, a great deal of research has been undertaken to propose effective techniques to resolve this reoccurring problem. The research presented in this paper is motivated by the fact that negative emotions can be caused by cyberbullying. This paper proposes cyberbullying detection models that are trained based on contextual, emotions and sentiment features. An Emotion Detection Model (EDM) was constructed using Twitter datasets that have been improved in terms of its annotations. Emotions and sentiment were extracted from cyberbullying datasets using EDM and lexicons based. Two cyberbullying datasets from Wikipedia and Twitter respectively were further improved by comprehensive annotation of emotion and sentiment features. The results show that anger, fear and guilt were the major emotions associated with cyberbullying. Subsequently, the extracted emotions were used as features in addition to contextual and sentiment features to train models for cyberbullying detection. The results demonstrate that using emotion features and sentiment has improved the performance of detecting cyberbullying by 0.5 to 0.6 recall. The proposed models also outperformed the state-of-the-art models by a 0.7 f1-score. The main contribution of this work is two-fold, which includes a comprehensive emotion annotated

dataset for cyberbullying detection, and an empirical proof of emotions as effective features for cyberbullying detection.

[7] "Adaptive Multimodal Emotion Detection Architecture for Social Robots," J. Heredia et al., in IEEE Access.

Emotion recognition is a strategy for social robots used to implement better Human-Robot Interaction and model their social behaviour. Since human emotions can be expressed in different ways (e.g., face, gesture, voice), multimodal approaches are useful to support the recognition process. However, although there exist studies dealing with multimodal emotion recognition for social robots, they still present limitations in the fusion process, dropping their performance if one or more modalities are not present or if modalities have different qualities. This is a common situation in social robotics, due to the high variety of the sensory capacities of robots; hence, more flexible multimodal models are needed. In this context, we propose an adaptive and flexible emotion recognition architecture able to work with multiple sources and modalities of information and manage different levels of data quality and missing data, to lead robots to better understand the mood of people in a given environment and accordingly adapt their behaviour. Each modality is analyzed independently to then aggregate the partial results with a previous proposed fusion method, called EmbraceNet+, which is adapted and integrated to our proposed framework. We also present an extensive review of state-of-the-art studies dealing with fusion methods for multimodal emotion recognition approaches. We evaluate the performance of our proposed architecture by performing different tests in which several modalities are combined to classify emotions using four categories (i.e., happiness, neutral, sadness, and anger). Results reveal that our approach is able to adapt to the quality and presence of modalities. Furthermore, results obtained are validated and compared with other similar proposals, obtaining competitive performance with state-of-the-art models.

[8] "Evaluation of Conversational Agents: Understanding Culture, Context and Environment in Emotion Detection," M. T. Teye, Y. M. Missah, E. Ahene and T. Frimpong, in IEEE Access.

Valuable decisions and highly prioritized analysis now depend on applications such as facial biometrics, social media photo tagging, and human robots interactions. However, the ability to successfully deploy such applications is based on their efficiencies on tested use cases taking into consideration possible edge cases. Over the years, lots of generalized solutions have

been implemented to mimic human emotions including sarcasm. However, factors such as geographical location or cultural difference have not been explored fully amidst its relevance in resolving ethical issues and improving conversational AI (Artificial Intelligence). In this paper, we seek to address the potential challenges in the usage of conversational AI within Black African society. We develop an emotion prediction model with accuracies ranging between 85% and 96%. Our model combines both speech and image data to detect the seven basic emotions with a focus on also identifying sarcasm. It uses 3-layers of the Convolutional Neural Network in addition to a new Audio-Frame Mean Expression (AFME) algorithm and focuses on model pre-processing and postprocessing stages. In the end, our proposed solution contributes to maintaining the credibility of an emotion recognition system in conversational AIs.

[9] "Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model," N. Aslam, F. Rustam, E. Lee, P. B. Washington and I. Ashraf, in IEEE Access.

The cryptocurrency market has been developed at an unprecedented speed over the past few years. Cryptocurrency works similar to standard currency, however, virtual payments are made for goods and services without the intervention of any central authority. Although cryptocurrency ensures legitimate and unique transactions by utilizing cryptographic methods, this industry is still in its inception and serious concerns have been raised about its use. Analysis of the sentiments about cryptocurrency is highly desirable to provide a holistic view of peoples' perceptions. In this regard, this study performs both sentiment analysis and emotion detection using the tweets related to the cryptocurrency which are widely used for predicting the market prices of cryptocurrency. For increasing the efficacy of the analysis, a deep learning ensemble model LSTM-GRU is proposed that combines two recurrent neural networks applications including long short term memory (LSTM) and gated recurrent unit (GRU). LSTM and GRU are stacked where the GRU is trained on the features extracted by LSTM. Utilizing term frequency-inverse document frequency, word2vec, and bag of words (BoW) features, several machine learning and deep learning approaches and a proposed ensemble model are investigated. Furthermore, TextBlob and Text2Emotion are studied for emotion analysis with the selected models. Comparatively, a larger number of people feel happy with the use of cryptocurrency, followed by fear and surprise emotions. Results suggest that the performance of machine learning models is comparatively better when BoW features are used.

The proposed LSTM-GRU ensemble shows an accuracy of 0.99 for sentiment analysis, and

0.92 for emotion prediction and outperforms both machine learning and state-of-the-art models.

[10] "Human–Robot Collaboration Using Sequential-Recurrent-Convolution-Network- Based Dynamic Face Emotion and Wireless Speech Command Recognitions," C. -L. Hwang, Y. -C. Deng and S. -E. Pu, in IEEE Access.

The proposed sequential recurrent convolution network (SRCN) includes two parts: one convolution neural network (CNN) and a sequence of long short-term memory (LSTM) models. The CNN is to achieve the feature vector of face emotion or speech command. Then, a sequence of LSTM models with the shared weight reflects a sequence of inputs provided by a (pre-trained) CNN with a sequence of input sub-images or spectrograms corresponding to face emotion and speech command, respectively. Simply put, one SRCN for dynamic face emotion recognition (SRCN-DFER) and another SRCN for wireless speech command recognition (SRCN-WSCR) are developed. The proposed approach not only effectively tackles the recognitions of dynamic mapping of face emotion and speech command with average generalized recognition rate of 98% and 96.7% but also prevents the overfitting problem in a noisy environment. The comparisons among mono and stereo visions, Deep CNN, and ResNet50 confirm the superiority of the proposed SRCN-DFER. The comparisons among SRCN-WSCR with noise-free data, SRCN-WSCR with noisy data, and multiclass support vector machine validate its robustness. Finally, the human-robot collaboration (HRC) using our developed omnidirectional service robot, including human and face detections, trajectory tracking by the previously designed adaptive stratified finite-time saturated control, face emotion and speech command recognitions, and music play, validates the effectiveness, feasibility, and robustness of the proposed method.

[11] "SCEP—A New Image Dimensional Emotion Recognition Model Based on Spatial and Channel-Wise Attention Mechanisms," B. Li, H. Ren, X. Jiang, F. Miao, F. Feng and
L. Jin, in IEEE Access.

Images are an important carrier for emotional expression. Human can understand emotions in image easily and quickly, whereas it is a very challenging task for machines to extract accurate emotions. In this study, we propose a novel spatial and channel-wise attention- based emotion prediction model, SCEP, to assist computers in recognizing the emotions of images more accurately. SCEP integrates both spatial attention and channel-wise weight mechanisms into a classical convolutional neural network (CNN) layer structure to predict

image emotions, on the grounds that the spatial attention mechanism can enhance the contrast between salient regions and potentially irrelevant regions, and that the channel-wise weight mechanism can emphasize informative features while suppressing less useful features. The SCEP model outputs emotion values in a continuous 2-D valence and arousal space, so that more emotions can be expressed than by simply discretely classifying emotions. To validate the effectiveness of our model, we use an existing image dataset with a widespread emotion distribution for testing. Extensive experiments show that when compared to base models (i.e. VGG and ResNet) without spatial attention or channel-wise mechanisms, SCEP can improve the accuracy of emotion prediction (evaluated by concordance correlation coefficient) by ∼3%- 5% in the arousal domain, and by ∼3-6% in the valence domain. Therefore, we conclude that using SCEP can bring higher accuracy in emotion prediction.

[12] "Context-Aware Emotion Recognition Based on Visual Relationship Detection," M. -H. Hoang, S. -H. Kim, H. -J. Yang and G. -S. Lee, in IEEE Access.

Emotion recognition, which is a part of affective computing, draws a lot of attention from researchers because of its broad applications. Unlike previous approaches with the aim to recognize humans' emotional state using facial expression, speech or gesture, some researchers see the potential of the contextual information from the scene. Hence, in addition to the employment of the main subject, the general background data is also considered as the complementary cues for emotion prediction. However, most of the existing works still have some limitations in deeply exploring the scene-level context. In this paper, to fully exploit the essences of context, we propose the emotional state prediction method based on visual relationship detection between the main target and the adjacent objects from the background. Specifically, we utilize both the spatial and semantic features of objects in the scene to calculate the influences of all context-related elements and their properties of impact (positive, negative, or neutral) on the main subject by a modified attention mechanism. After that, the model incorporates those features with scene context and body features of the target person to predict their emotional states. Our experimental results achieve state-of-the-art performance on the CAER-S dataset and competitive results on the EMOTIC benchmark.

[13] "Utilizing External Knowledge to Enhance Semantics in Emotion Detection in Conversation," F. Ren and T. She, in IEEE Access.

Enabling machines to emotion recognition in conversation is challenging, mainly because the information in human dialogue innately conveys emotions by long-term

experience, abundant knowledge, context, and the intricate patterns between the affective states. We address the task of emotion recognition in conversations using external knowledge to enhance semantics. We propose KES model, a new framework that incorporates different elements of external knowledge and conversational semantic role labeling, where build upon them to learn interactions between interlocutors participating in a conversation. We design a self-attention layer specialized for enhanced semantic text features with external commonsense knowledge. Then, two different networks composed of LSTM are responsible for tracking individual internal state and context external state. In addition, the proposed model has experimented on three datasets in emotion detection in conversation. The experimental results show that our model outperforms the state-of-the-art approaches on most of the tested datasets.

[14] "Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features," N. Samadiani, G. Huang, Y. Hu and X. Li, in IEEE Access.

Facial expressions have been proven to be the most effective way for the brain to recognize human emotions in a variety of contexts. With the exponentially increasing research for emotion detection in recent years, facial expression recognition has become an attractive, hot research topic to identify various basic emotions. Happy emotion is one of such basic emotions with many applications, which is more likely recognized by facial expressions than other emotion measurement instruments (e.g., audio/speech, textual and physiological sensing). Nowadays, most methods have been developed for identifying multiple types of emotions, which aim to achieve the best overall precision for all emotions; it is hard for them to optimize the recognition accuracy for single emotion (e.g., happiness). Only a few methods are designed to recognize single happy emotion captured in the unconstrained videos; however, their limitations lie in that the processing of severe head pose variations has not been considered, and the accuracy is still not satisfied. In this paper, we propose a Happy Emotion Recognition model using the 3D hybrid deep and distance features (Happy ER-DDF) method to improve the accuracy by utilizing and extracting two different types of deep visual features. First, we employ a hybrid 3D Inception-Res Net neural network and long-short term memory (LSTM) to extract dynamic spatial-temporal features among sequential frames. Second, we detect facial landmarks' features and calculate the distance between each facial landmark and a reference point on the face (e.g., nose peak) to capture their changes when a person starts to smile (or laugh).

We implement the experiments using both feature-level and decision-level fusion techniques on three unconstrained video datasets. The results demonstrate that our Happy ER-DDF method is arguably more accurate than several currently available facial expression models.

2.4 Summary

The paper "Three Way Emotion Detector and Recommendation" presents a ground breaking multi-modal emotion detection platform that integrates facial expressions, voice tonality, and text sentiment analysis. It outlines the methodology, design, and potential applications of this system, emphasizing its significance in human-computer interaction, virtual assistant technology, and mental health support. Through detailed discussions on data collection, preprocessing, module design, and integration strategies, the paper highlights the comprehensive development process. Moreover, it showcases the system's validation against diverse datasets, benchmarking efforts, and user evaluations, solidifying its effectiveness and superiority over single-modal approaches. Overall, the paper represents a significant advancement in multi-modal emotion detection, offering promising implications across various domains.

Recent advancements in emotion detection (ED) research, particularly within the realm of online social media (OSM) platforms like Twitter, have prompted innovative methodologies to capture and analyze users' emotional states. One approach introduces a novel feature representation strategy for Twitter data, leveraging Genetic Algorithm (GA) for constructing

input features and a voting ensemble classifier for improved emotion classification accuracy.

Another study delves into emotion recognition across diverse scenes, integrating facial, body pose, and contextual information to surpass existing methods in understanding environmental influences on human emotions. Meanwhile, researchers investigate fine-grained emotion analysis to enhance the detection of implicit hate speech, underscoring the significance of nuanced emotional features in mitigating hate-based activities online. In a distinct domain, a novel FF-BTP model tailored for crowd sentiment analysis through sound-based community emotion detection demonstrates high accuracy, promising efficient emotion detection with reduced computational requirements. Emotion recognition in Human-Computer Interface (HCI) applications receives attention with the development of a novel MEWODL-ER technique, leveraging Modified Earthworm Optimization and Deep Learning for improved emotion classification accuracy.

Additionally, comprehensive emotion annotation is utilized to enhance cyberbullying detection models, showcasing the effectiveness of emotion features alongside contextual and sentiment features. Emotion detection in social robots sees advancements with an adaptive multimodal architecture, enabling robust recognition of dynamic face emotions and wireless speech commands for effective human-robot collaboration. Other studies focus on image-based emotion recognition, utilizing spatial and channel-wise attention mechanisms to enhance emotion prediction accuracy, and integrating external knowledge for improved emotion detection in conversations. Lastly, a novel method for recognizing happy emotions in unconstrained videos demonstrates superior accuracy through the utilization of 3D hybrid deep and distance features, addressing challenges such as severe head pose variations. In summary, these diverse methodologies highlight the importance of contextual understanding and nuanced feature extraction for accurate emotion recognition across various domains.

In planning phase study of reliable and effective algorithms is done. On the other hand data were collected and were preprocessed for more fine and accurate results. Since huge amount of data were needed for better accuracy, we have collected the data surfing the internet. Since, we are new to this project we have decided to use local binary pattern algorithm for feature extraction and support vector machine for training the dataset. We have decided to implement these algorithms by using Open CV framework.

3.1 Analysis and Feasibility Study

Modalities Integration Feasibility: The feasibility of integrating three distinct modalities facial expressions analysis, voice characteristics assessment, and text sentiment analysis is assessed. This involves evaluating the compatibility and interoperability of different technologies, such as Convolutional Neural Networks (CNNs), librosa library for audio processing, and Natural Language Processing (NLP) techniques.

Technological Viability: The viability of employing CNNs for facial expression recognition, utilizing Mel-Frequency Cepstral Coefficients (MFCC) for voice data analysis, and applying NLP and machine learning techniques for textual sentiment analysis is examined. This includes assessing the availability of relevant libraries, tools, and datasets for each modality.

Emotion Detection Accuracy: The accuracy and reliability of emotion detection using each modality individually and in combination are investigated. This involves conducting preliminary experiments and feasibility tests to determine the effectiveness of Convolutional Neural Networks (CNNs), MFCC concepts, and NLP algorithms in accurately detecting and interpreting emotions from facial expressions, voice characteristics, and textual content.

Recommendation System Viability: The feasibility of implementing a recommendation system based on detected emotions is explored. This includes assessing the availability of relevant data sources for generating personalized recommendations, such as healthcare professional databases, movie and podcast libraries, news articles, and YouTube video recommendations.

User Acceptance and Ethical Considerations: The potential acceptance of the proposed emotion detection and recommendation system by users is evaluated. Additionally, ethical considerations related to user privacy, data security, and the responsible use of emotional data are addressed to ensure the system's ethical integrity and compliance with regulations.

3.2 Requirement Specification:

Requirement analysis is mainly categorized into two types:

1. Functional requirements:

The functional requirements for a system describe what the system should do. Those requirements depend on the type of software being developed, how the system should react to particular inputs and how the system should behave in particular situation.

2. Non-Functional requirements:

Non-functional requirements are requirements that are not directly concerned with the specified function delivered by the system. They may relate to emergent system properties such as reliability, response time and store occupancy. Some of the nonfunctional requirements related with this system are here by below:

a) Reliability

Reliability based on this system defines the evaluation result of the system, correct identification of the facial expressions and maximum evaluation rate of the facial expression recognition of anyinput images.

b) Ease of Use

The system is simple, user friendly, graphics user interface implemented so any can use this system without any difficulties.

c) Feasibility Study

Before starting the project, feasibility study is carried out to measure the viable of the system. Feasibility study is necessary to determine if creating a new or improved system is friendly with the cost, benefits, operation, technology, and time.

d) Technical Feasibility

Technical feasibility is one of the first studies that must be conducted after the project has been identified. Technical feasibility study includes the hardware and software devices. The required

technologies (C++ language and CLion IDE) existed.

e) Operational Feasibility

Operational Feasibility is a measure of how well a proposed system solves the problem and takes advantage of the opportunities identified during scope definition. The following points were considered for the project's technical feasibility:

- The system will detect and capture the image of face
- The captured image is then analyzed based on the category.

f) Economic Feasibility

The purpose of economic feasibility is to determine the positive economic benefits that include quantification and identification.

3.3 Hardware Requirement Specification

- Processor: Any Processor above 3 GHZ
- Ram: 4 GB
- Hard Disk: 10 Gb.
- Compact Disk: 650 Mb.
- Input device: Standard Keyboard and Mouse.
- Output device: VGA and High-Resolution Monitor.

3.4 Software Requirement Specification

- Operating System: Windows XP/Above
- Software Tool: Open CV Python
- Coding Language: Python

Open CV:

Open CV is a versatile library of programming functions designed for real-time computer vision asks. It provides a modular structure with various shared or static libraries.

The image processing module of Open CV offers a range of capabilities such as linear and non- linear image filtering, geometric transformations, color space conversion, histograms, and more.

Our project utilizes specific libraries within Open CV, including Viola-Jones or Haar classifier, LBPH (Lower Binary Pattern Histogram) face recognizer, and Histogram of Oriented Gradients (HOG). Open CV is primarily written in C++, with interfaces available in C++, Python, Java, and MATLAB. It is compatible with different platforms such as Windows, Linux, mac OS, as well as mobile platforms like Android, iOS, and Blackberry.

Open CV finds application in various fields such as facial recognition, gesture recognition, object identification, mobile robotics, and image segmentation. Our project utilizes Open CV version 2 and employs its functionalities for gesture-controlled camera access, image capture, image-to-text conversion, and voice conversion.

Fig 3.1 Open CV

The purpose of image processing can be categorized into five groups:

1. Visualization: Enhancing the visibility of objects that are not easily observable.
2. Image sharpening and restoration: Improving the quality and clarity of images.
3. Image retrieval: Searching for specific images based on predefined criteria.
4. Measurement of patterns: Extracting measurements and features from images.
5. Image recognition: Identifying and classifying objects within images.

eSpeak :

It is a compact open-source software speech synthesizer for English and 11other languages for Linux and Windows platform. It is used to convert text to voice. It supports many languages in a small size. The programming for espeak software is done using rule files with feedback. It supports SSML. It can be modified by voice variant. These are text files which can change1characteristics such as1pitchrange, add effects1such as echo, whisper and croaky voice, or make systematic adjustments to formant frequencies to change the sound of the voice. The default speaking speed of 180 words per minute is too fast to be intelligible. In our project Espeak is used to convert the text to voice signal.

Fig 3.2 e-Speak.

SYSTEM DESIGN

4.1 INPUT/ OUTPUT DESIGN:

System design shows the overall design of system. In this section we discuss in detail the design aspects of the system.

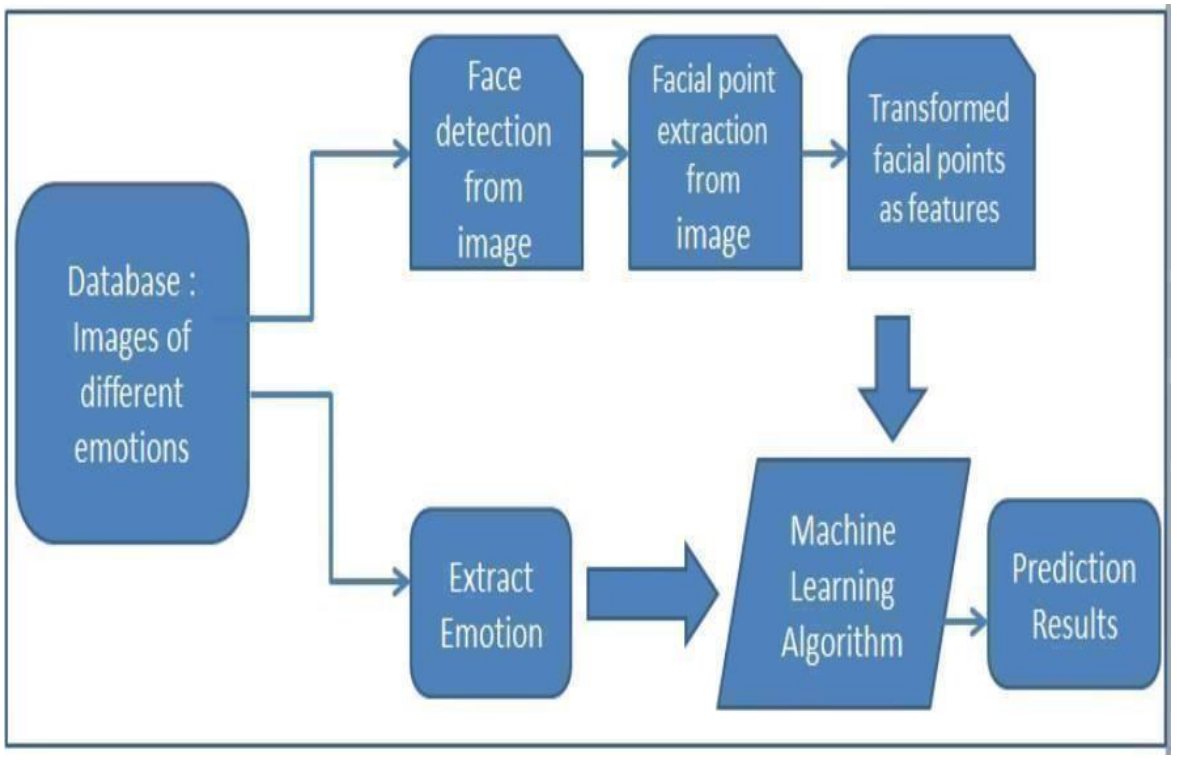


Fig: 4.1.1 Facial Emotion Recognition Using Machine Learning.

4.1.1 Phases In Facial Expression Recognition

The facial expression recognition system is trained using a supervised learning approach, whereit utilizes images of various facial expressions. The system consists of a training phase and a testing phase, which involves several stages including image acquisition, face detection, image preprocessing, feature extraction, and classification. The process can be summarized as follows:

1. Image Acquisition.
2. Face detection.

1. **Image Acquisition**

Images containing facial expressions are collected either from a dataset or through real-time capture using a camera.

2. **Face detection**

- **Image Pre-processing:**

Image pre-processing includes the removal of noise and normalization against the variation of pixel position or brightness.

a) Color Normalization.

b) Histogram Normalization.

- **Feature Extraction**

Selection of the feature vector is the most important part in a pattern classification problem. The image of face after pre-processing is then used for extracting the important features. The inherent problems related to image classification include the scale, pose translation and variations in illumination level.

4.2 PATTERN

The important features are extracted using LBP algorithm which is described below:

Local Binary Pattern (LBP) :

LBP is a feature extraction technique that encodes the local structure of each pixel in an image. It compares each pixel with its eight neighbors in a 3x3 neighborhood by subtracting the center pixel value. Negative values are encoded as 0 and others as 1. The binary values are merged in a clockwise direction, starting from the top-left neighbor, to form a binary number. This binary number is then converted to decimal and used to label the given pixel. These derived binary numbers are known as LBP codes.

4.3 Convolution Neural Networks:

CNNs are vital in Deep Learning for Computer Vision tasks. They analyze and extract features from images using convolutional layers, pooling layers, and fully connected layers. This enables them to learn hierarchical representations of visual data, allowing them to classify images, detect objects, and segment images. CNNs revolutionized Computer Vision by capturing spatial dependencies and hierarchical patterns, leading to remarkable success in diverse applications.

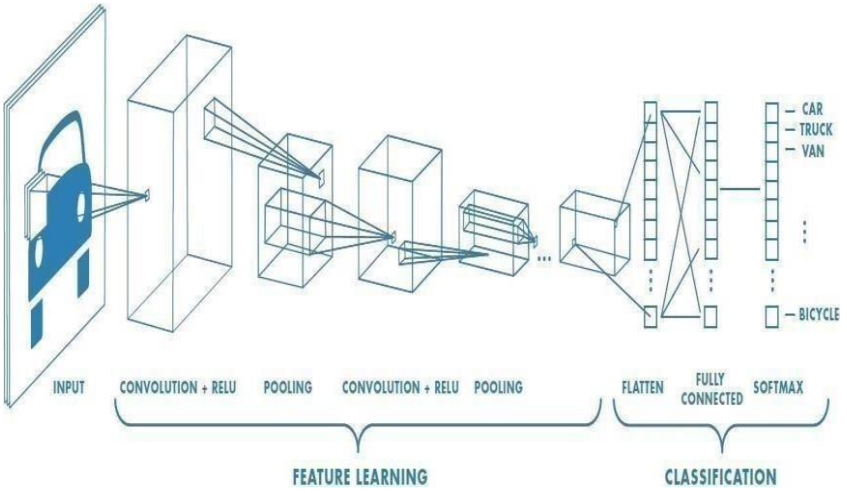


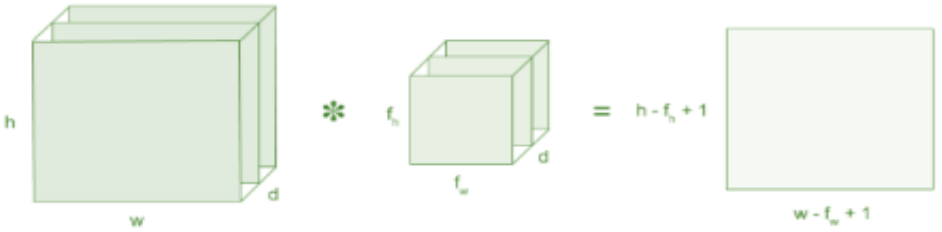
Fig: 4.3.1 Neural network with many convolutional layers

4.4 Convolution Layer:

Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel.

Fig: 4.4.1 Image matrix multiplies kernel or filter matrix(1).

- An image matrix (volume) of dimension **(h x w x d)**
- A filter **(f_h x f_w x d)**
- Outputs a volume dimension **(h - f_h + 1) x (w - f_w + 1) x 1**



Consider a 5 x 5 whose image pixel values are 0, 1 and filter matrix 3 x 3 as shown in below

Fig 4.4.2: Image matrix multiplies kernel or filter matrix(2).

Then the convolution of 5 x 5 image matrix multiplies with 3 x 3 filter matrix which is called **"Feature Map"** as output shown in below

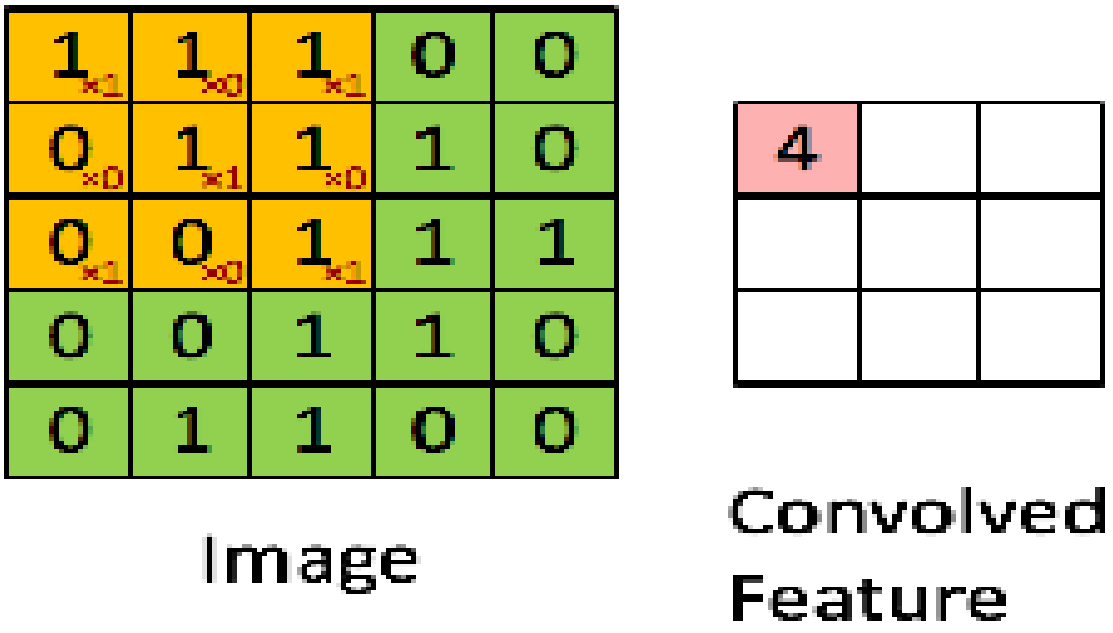


Fig 4.4.3: 3 x 3 Output matrix

Convolution of an image with different filters allows for operations like edge detection, blurring, and sharpening. By applying various types of filters (kernels), we can obtain different convolutional images that represent these operations.

Strides

In convolution, the stride refers to the number of pixel shifts applied to the input matrix. A stride of 1 means the filters move one pixel at a time, while a stride of 2 means the filters move two pixels at a time, and so on. When using a stride of 2, the filters skip every other pixel during the convolution operation. The figure below illustrates how convolution works with a stride of 2..

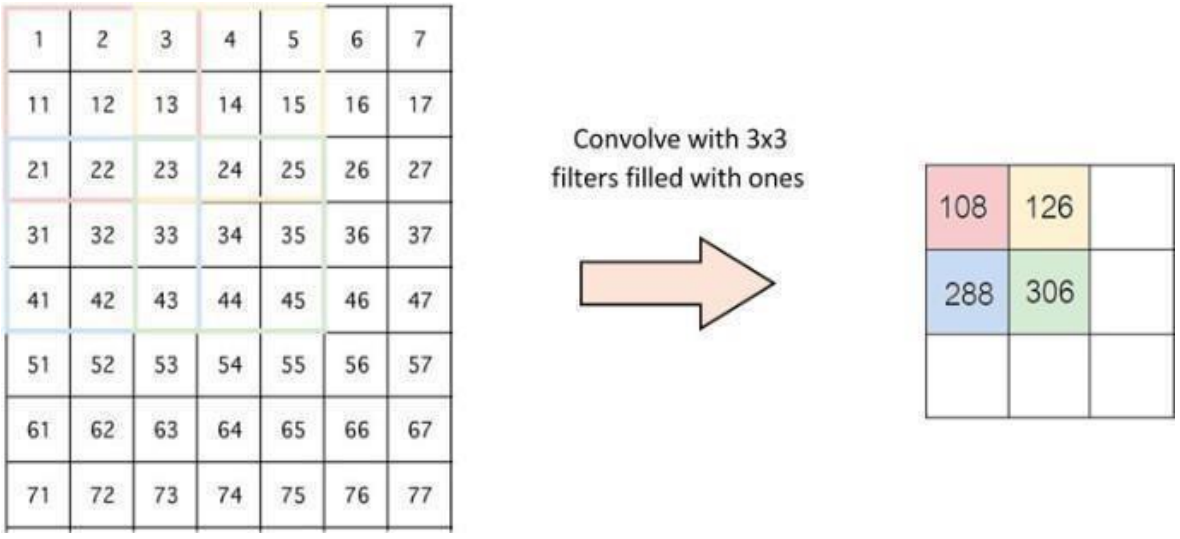


Fig 4.4.4: Stride of 2 pixels.

In valid padding, the part of the image where the filter did not fit is dropped, keeping only the valid portion of the image. This means that the filter is applied only to the parts of the image where it fully overlaps without exceeding the boundaries. Valid padding ensures that the output size is smaller than the input size, as only the parts of the image that can be fully convolved with the filter are considered.

Padding

In some cases, the filter may not perfectly fit the input image. In such situations, we have two options:

1. **Zero-padding:**

   This involves padding the input image with zeros (zeros-padding) so that the filter can fit properly. Padding adds additional rows and columns of zeros aroundthe input image, ensuring that the filter covers all the pixels. This is a commonly used technique to maintain the spatial dimensions of the input and output.

2. **Remove the excess:**

Alternatively, we can choose to remove the parts of the image where the filter does not fit. This means discarding the pixels that lie outside the filter's coverage area. This approach reduces the size of the output compared to the input image.

Both options have their advantages and depend on the specific requirements of the task at hand.

Non Linearity (ReLU) :

ReLU stands for Rectified Linear Unit, which is a non-linear activation function commonly usedin Convolutional Neural Networks (CNNs). It applies the function $f(x) = \max(0, x)$, where x represents the input.

ReLU is important in CNNs because it introduces non-linearity into the network, allowing it to learn and model complex relationships in the data. By applying ReLU, the network becomes capable of learning non-negative linear values, which is often suitable for real-world data. This non-linearity is crucial for CNNs to capture and represent more intricate patterns and features in images, leading to improved performance in tasks such as image classification, object detection, and image segmentation.

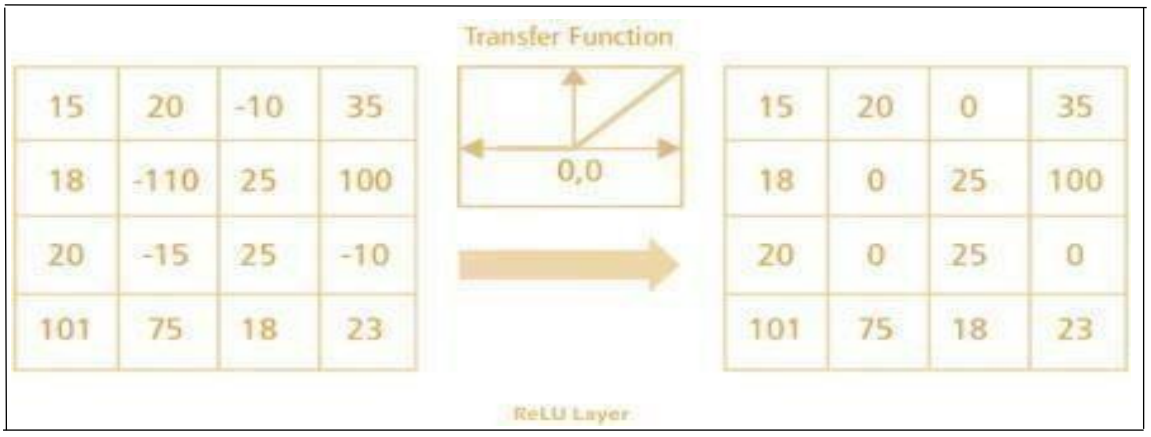


Fig 4.4.5: ReLU operation.

ReLU, tan h, and sigmoid are popular nonlinear activation functions in deep learning. While tan h and sigmoid functions can be used, ReLU is preferred by data scientists due to its computational efficiency and ability to mitigate the vanishing gradient problem. ReLU's sparsity property also aids interpretability and efficiency.

Pooling Layer :

Pooling layers are utilized in Convolutional Neural Networks (CNNs) to reduce the number of parameters when working with large images. This technique, also known as subsampling or down sampling, helps in decreasing the dimensionality of each feature map while retaining crucial information. There are several types of spatial pooling, including max pooling, average pooling, and sum pooling.

In max pooling, the largest element from the rectified feature map is selected, while average pooling takes the average of the elements in the feature map. Sum pooling, on the other hand, calculates the sum of all elements in the feature map. These pooling operations aid in reducing the spatial dimensions of the feature maps, enabling the network to focus on the most significant features while discarding less important details.

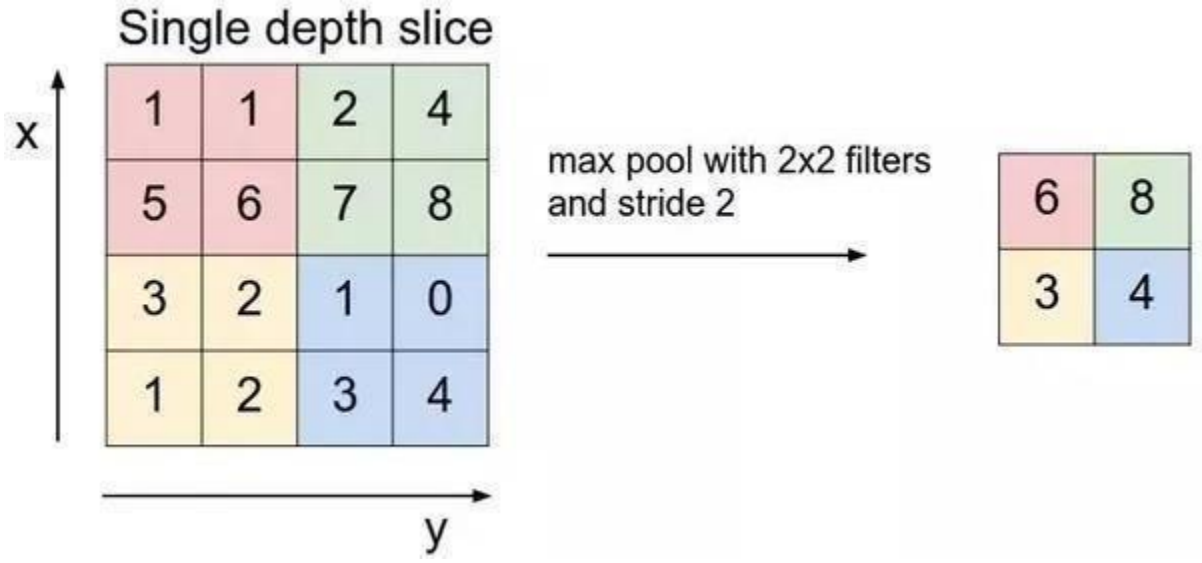


Fig 4.4.6: Max Pooling.

Fully Connected Layer:

The fully connected (FC) layer in a neural network refers to a layer where the input matrix is flattened into a vector and then fed into the layer. This layer is similar to a traditional neural network layer, where each neuron is connected to every neuron in the previous and following layers. The FC layer enables the network to learn complex patterns and relationships by considering the interactions between all the features in the flattened vector.

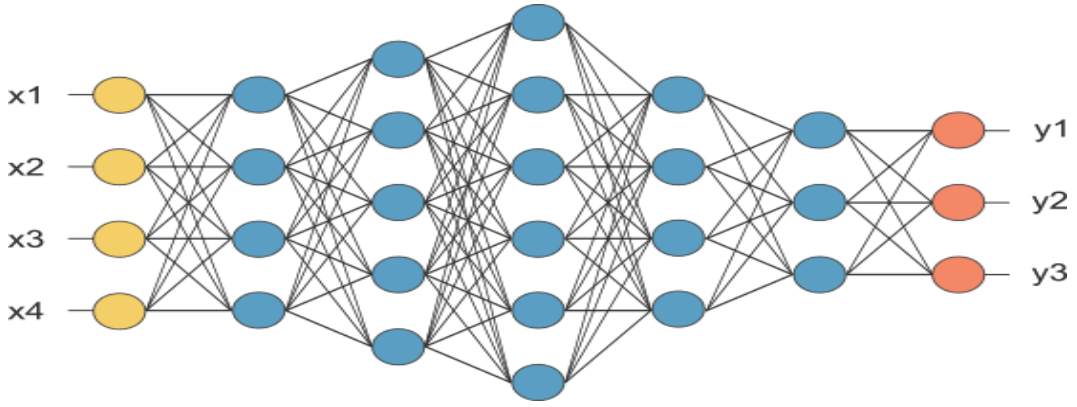


Fig 4.4.7: After pooling layer, flattened as FC layer

In the above diagram, the feature map matrix will be converted as vector (x1, x2, x3, …). With the fully connected layers, we combined these features together to create a model. Finally, we have an activation function such as Soft Max or sigmoid to classify the outputs as cat, dog, car.

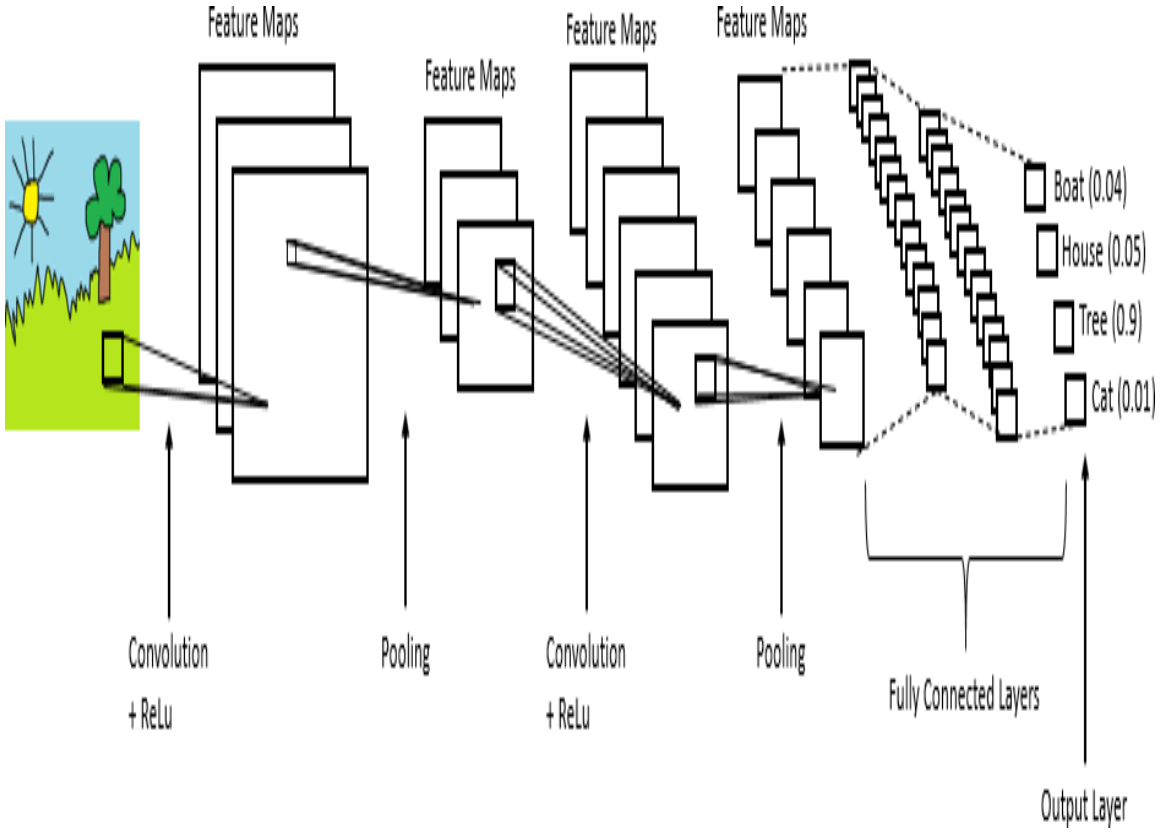


Fig 4.4.8: Complete CNN architect

Implementation and Testing

Implementation and testing are two crucial steps in software development. In implementation, the software is coded and built according to the requirements and specifications. In testing, the software is tested for errors, bugs, and other issues that could affect its functionality and performance. Here are some steps you can follow for implementation and testing:

Implementation: The implementation process involves several steps, including coding, integration, and deployment.

- Understand the requirements and specifications thoroughly.
- Design the software architecture, database schema, and user interfaces.
- Write the code according to the design and specifications.
- Test each module or component of the software to ensure it functions as expected.
- Integrate all the modules and components to form a complete software system
- Test the software system as a whole to ensure it meets the requirements and specifications.
- Document the code and system architecture for future maintenance and enhancement.

Testing: Testing is an essential part of software development that ensures that the software meets the project's requirements and functions correctly.

- Create a test plan that covers all the features and functions of the software.
- Write test cases for each feature and function.
- Conduct unit testing, integration testing, system testing, and acceptance testing to ensure thesoftware meets all the requirements and specifications.
- Identify and report any issues or bugs found during test.

It is important to conduct both implementation and testing with due diligence and thoroughness to ensure the software functions optimally and meets the requirements and

expectations of the end-users.

5.1 General Implementation Discussions

5.1.1 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a type of deep neural network specifically designed to analyze visual data. It's primarily used for tasks such as image recognition, object detection, and image classification, but it can also be applied to other types of data with spatial relationships, such as time-series data in signal processing or spatial data in natural language processing. CNNs are composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply convolution operations to the input data, which involve sliding a set of learnable filters (kernels) over the input to extract spatial features. The pooling layers then downsample the feature maps obtained from the convolutional layers to reduce the computational complexity and control over fitting. Finally, fully connected layers are used to classify the features extracted by the previous layers into different classes.

Convolutional Neural Networks (CNNs) consist of convolutional layers, pooling layers, and fully connected layers. These layers extract features hierarchically from input data and learn representations suitable for the task at hand. The model's architecture is defined by stacking these layers, typically starting with convolutional layers that extract features from the input images, followed by pooling layers to reduce spatial dimensions and fully connected layers for classification. Implementing a Convolutional Neural Network (CNN) involves defining its architecture with convolutional, pooling, and fully connected layers, compiling it with an optimizer and loss function, and training it on labeled data. During training, the model learns to extract features and classify inputs by adjusting its weights through backpropagation. Techniques like dropout and regularization are used to prevent overfitting. After training, the model is evaluated using separate test data to assess its performance metrics like accuracy. Once validated, the trained CNN can be deployed for real-world applications such as image classification or object detection.

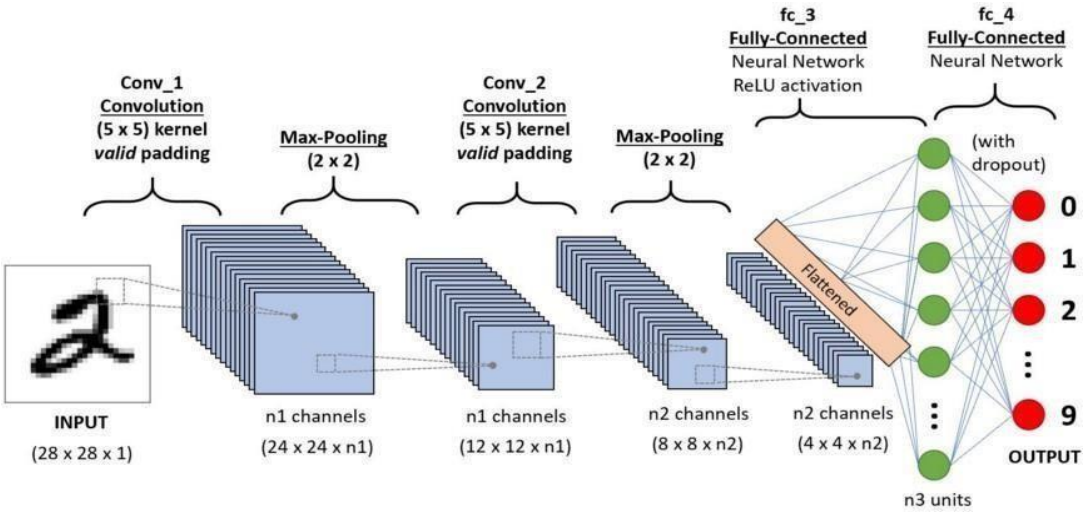


Fig 5.1: Working of CNN.

5.1.2 Natural language processing

Natural Language Processing (NLP) involves computational techniques to understand, interpret, and generate human language. It encompasses tasks like text classification, sentiment analysis, machine translation, and chatbot development. NLP techniques enable machines to comprehend and process human language, enabling applications such as virtual assistants, search engines, and language translation services. Key components include tokenization, part-of-speech tagging, named entity recognition, and syntactic parsing. Techniques like word embeddings and deep learning models have significantly advanced NLP capabilities in recent years, allowing for more accurate and nuanced language understanding. NLP finds applications across various domains, including healthcare, finance, customer service, and education, revolutionizing how humans interact with technology and data.

Implementing multi-class text classification in NLP involves gathering a dataset of text documents and their corresponding labels, followed by preprocessing to remove noise and normalize the text. Next, the text data is transformed into numerical vectors using techniques like Bag-of-Words or word embeddings. A suitable classification model, such as logistic regression, Naive Bayes, SVM, or deep learning models like RNNs or Transformers, is selected for the task. The model is trained on a portion of the dataset while monitoring performance on a validation set to prevent overfitting. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed on a separate test set to assess the model's performance. Fine-tuning and optimization are conducted by experimenting with different

architectures and hyperparameters to enhance performance. Finally, the deployed model can be integrated into production systems or deployed on cloud platforms after thorough testing and validation. Throughout the implementation process, ethical considerations such as bias mitigation and fairness in dataset representation are crucial aspects to address.

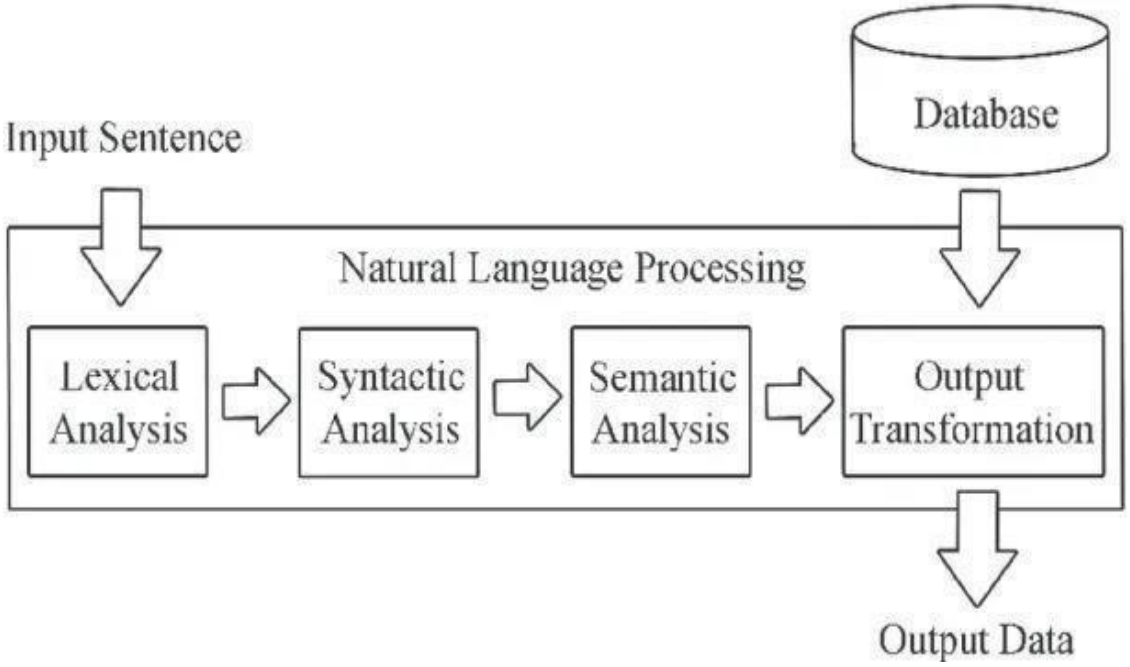


Fig 5.2: Working of CNN.

5.1.3 Librosa

Librosa is a Python package for music and audio analysis. It provides functionalities for extracting features from audio signals, such as spectrograms, mel-frequency cepstral coefficients (MFCCs), chromagrams, and tempo. The package allows users to load audio files in various formats, manipulate audio signals, and perform signal processing operations like resampling and filtering. Librosa also offers tools for visualizing audio data and analyzing music properties like beats, onsets, and harmonic-percussive separation. Its straightforward interface and extensive documentation make it a popular choice for researchers, audio engineers, and music enthusiasts seeking to explore and analyze audio content efficiently within the Python ecosystem.

5.2 Test Setup

Librosa is a Python package for music and audio analysis. It provides functionalities for extracting features from audio signals, such as spectrograms, mel-frequency cepstral coefficients (MFCCs), Chroma grams, and tempo. The package allows users to load audio files in various formats, manipulate audio signals, and perform signal processing operations like Testing is the process of evaluating a system or its component(s) with the intent to find whetherit satisfies the specified requirements or not. Testing is executing a system in order to identify any gaps, errors, or missing requirements in contrary to the actual requirements.

5.2.1 Unit testing

Unit testing is a software development process in which the smallest testable parts of an application, called units, are individually and independently scrutinized for proper operation. Unit testing is often automated but it can also be done manually. The goal of unit testing is to isolate each part of the program and show that individual parts are correct in terms of requirements and functionality. Test cases and results are shown in the Tables.

GIVEN INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
Input image	Should Capture the image	Image Capture Success full	PASS
Tested for Different images	Created blob For HumanFace	Face Detection Successful	PASS
Different images beingtested	Should detect facial landmarks	Same as Expected	PASS

Table 5.1: Unit testing Test Cases.

5.2.2 Integration testing

Integration testing is a level of software testing where individual units are combined and tested as a group. The purpose of this level of testing is to expose faults in the interaction between integrated units. Test drivers and test stubs are used to assist in Integration Testing. Integration testing is defined as the testing of combined parts of an application to determine if they function correctly. It occurs after unit testing and before validation testing. Integration testing can be done in two ways: Bottom-up integration testing and Top-down integration testing.

1. **Bottom-up Integration**

This testing begins with unit testing, followed by tests of progressively higher-level combinations of units called modules or builds.

2. **Top-down Integration**

In this testing, the highest-level modules are tested first and progressively, lower-level modules are tested thereafter.

In a comprehensive software development environment, bottom-up testing is usually done first, followed by top-down testing. The process concludes with multiple tests of the complete application, preferably in scenarios designed to mimic actual situations. Table below shows the test cases for integration testing and their results.

GIVEN INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
Click and select image	Should create blob for Face	Image captured and Face Detected	PASS
Image input	Face Recognition andLandmark detection	Facial landmarkscalculated	PASS

Table 5.2: Integration testing Test Cases.

5.2.3 System testing

System testing of software or hardware is testing conducted on a complete, integrated system to evaluate the system's compliance with its specified requirements. System testing falls within the scope of black-box testing, and as such, should require no knowledge of the inner design of the code or logic.

GIVEN INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
Input image	Should detect behavior and update	Behavior Updated Successfully	PASS

Table 5.3: System testing Test Cases.

5.2.4 Acceptance testing

Acceptance testing is actually a series of different tests whose primary purpose is to fullyexercise the computer-based system. Include recovery testing crashes, security testing for unauthorized.

GIVEN INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
Run Applications Windows XP , Windows 10	Functionality should be according to given criteria	Working as expected output.	PASS

Table 5.4: Acceptance testing Test Cases.

5.2.5 Validation testing

At the culmination of integration testing, software is completely assembled as a packages; interfacing errors have been covered and corrected, and final series of software tests

-validating testing may begin. Validation can be defined in many ways, but a simple definition is that validation succeeds when software functions in a manner that can be reasonably expected by customers. Reasonable expectation is defined in the software requirement

specification- a document that describes all users' visible attributes of the software.

Implementation code

The Fig 5.1.1, show the implementation of the code that id designed in python. It corresponds to the functionalities provided by the image input working system.

```
def predict_mood():
    final_label = None
    cap = cv.VideoCapture(0)
    got = False
    while True:
        ret,frame = cap.read()
        gray = cv.cvtColor(frame,cv.COLOR_BGR2GRAY)
        faces = face_classifier.detectMultiScale(gray,1.3,5)
                for x,y,w,h in faces:
            cv.rectangle(frame,(x,y),(x+w,y+h),(255,0,0),2)
            roi_gray = gray[y:y+h,x:x+w]
            roi_gray = cv.resize(roi_gray,(48,48),interpolation=cv.INTER_AREA)
                if(np.sum([roi_gray])!=0):
                roi = roi_gray.astype('float')/255.0
                roi = img_to_array(roi)
                roi = np.expand_dims(roi,axis=0)
                preds = model12.predict(roi)[0]
                label = classes12[preds.argmax()]
                label_position = (x,y)
                final_label = label
                cv.putText(frame,label,label_position,cv.FONT_HERSHEY_COMPLEX,
2,(0,255,0))
            else:
                cv.putText(frame,'No Face
Found',(20,60),cv.FONT_HERSHEY_COMPLEX,2,(0,0,255))
        cv.imshow('Emotion Detector',frame)
        if cv.waitKey(1) & 0xFF == ord('q'):
            break
    cap.release()
    cv.destroyAllWindows()
    # print("Done")
    print(final_label)
```

Fig 5.2.1 Camera input code

The Fig 5.1.2, show the implementation of the code that id designed in python. It corresponds to the functionalities provided by the audio input working system.

```
def audio():
    fs = 44100
    seconds = 5.0
    myrecording = sd.rec(int(seconds * fs), samplerate=fs, channels=2)
    sd.wait()
    write('output10.wav', fs, myrecording)
    X, sample_rate = librosa.load('output10.wav')
    sample_rate = np.array(sample_rate)
    mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate,
n_mfcc=40).T,axis=0)
    livedf2 = mfccs
    livedf2= pd.DataFrame(data=livedf2)
    livedf2 = livedf2.stack().to_frame()
    x = np.expand_dims(livedf2, axis=2)
    x = np.expand_dims(x, axis=0)
   # predictions = loaded_model.predict_classes(x)
    predictions = np.argmax(loaded_model.predict(x), axis=-1)
    # covert emotion to voive
    mood = convert_class_to_emotion(predictions)
    print(mood)
```

Fig 5.2.2 Audio input code

The Fig 5.1.3, show the implementation of the code that id designed in python. It corresponds to the functionalities provided by the text input working system.

```
def mood_text():
    form=SubmitText()
    if form.validate_on_submit():
        text = str(form.text.data)
        predictions = []
        predicted_labels = []
        j = 0
        emotions = ["anger", "disgust", "fear", "joy", "neutral", "sadness", "shame",
"surprise"]
        pred = list(pipe_lr.predict_proba([text])[0])
        preds = []
        for j in pred:
            ii = int(j*100)
            preds.append(ii)
        predictions.append(preds)
        predicted_labels.append(emotions)
        final_preds = []
        final_labels = []
        final_preds.append(max(preds))
```

```
        final_labels.append(emotions[preds.index(max(preds))])
        plt.bar(emotions, pred)
        plt.title(text)
        plt.savefig('emotions/static/results/graph.jpg')
        print("text entered is \n\n\n\n {} \n\n\n\n ".format(text))
        check=['joy', 'neutral','surprise']
        out = emotions[preds.index(max(preds))]
        print('text emotion is', out)
        if out =="joy" :
            return render_template("happy.html")
        elif out =="neutral":
            return render_template("Neutral.html")
        elif out =="surprise" :
            return render_template("surprise.html")
        else :
            return render_template("sad.html")

    return render_template('text_page.html',title='Submit Text',form=form)
```

Fig 5.2.3 Text input code.

The fig 5.1.4 show the display of the designing of the user interface which allows the user to experience the working of the system.

```
function Songs_Eng() {
            window.open("https://youtu.be/yM594a-1wDA?si=tJvnZRMnPPa5Ld_i");
        }
        function Songs_tamil() {
            window.open("https://open.spotify.com/playlist/7jT0GOgNIBB2VNDrNBf
gXU");
        }
        function Songs_kannada() {
            window.open("https://open.spotify.com/playlist/4KRZnq8y3FghxBZcq6Q
Dzq");
         }
        function Songs_telegu() {
            window.open("https://open.spotify.com/playlist/37i9dQZF1DWUEWjDsV7
AgX");
         }
```

Fig 5.2.4 User interface code

Results and Discussions

The working of emotion based music player is demonstrated in the following content with the helpof snapshots.

6.1 Login Page

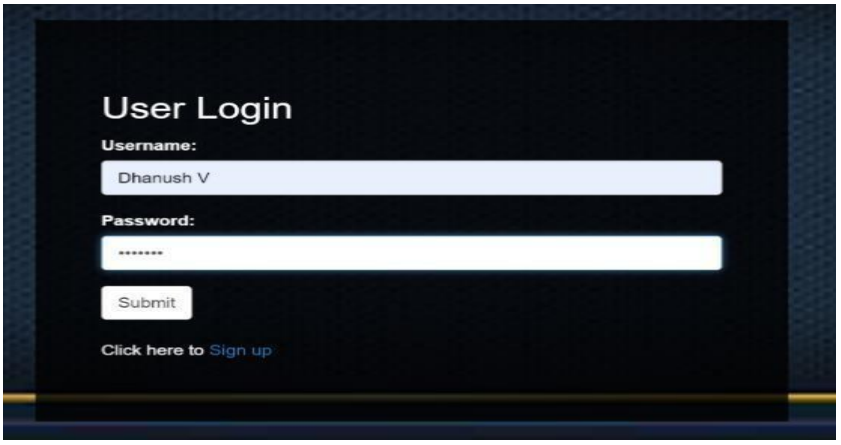


Fig 6.1: Login Page

The above figure shows the login page. The user has to enter the website by the entering the login information.

6.2 User Interface

Fig 6.2: User Interface.

The above figure shows the interface that will be displayed to the user on entering the application.It provides three entry prompts to the user. The first on is the camera which allow users to record the facial emotion. The second is recorder which is used to record user's voice to detect the emotion. The third is the text which is used to detect the emotion of a person using the text as input.

6.3 Text- Emotion Detection

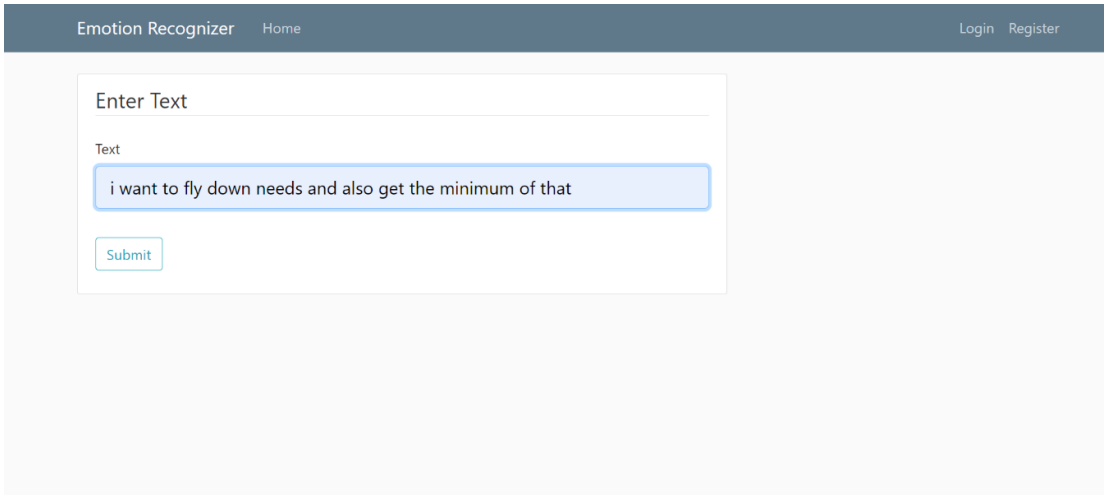


Fig 6.3: Text-Based Emotion Detection.

The above figure shows the emotion detection of a person and displays the emotion of the personwith the help of textual input of the user and detects the user emotion.

6.4 Facial Emotion Detection

The below figure shows the facial detection of a person and displays the emotion of the person.

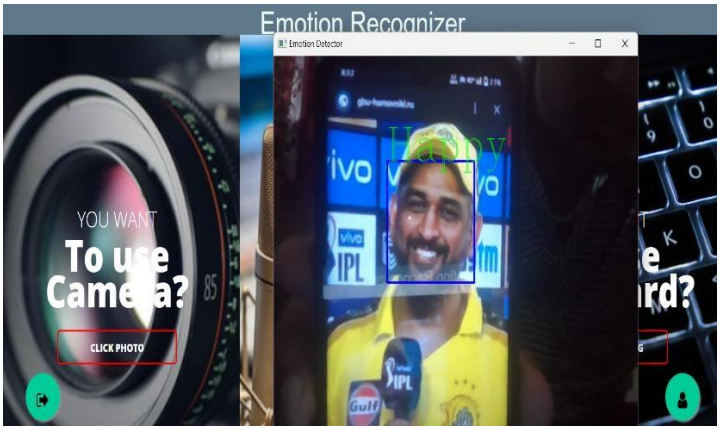


Fig 6.4: device camera captures person's face and detects the emotion .

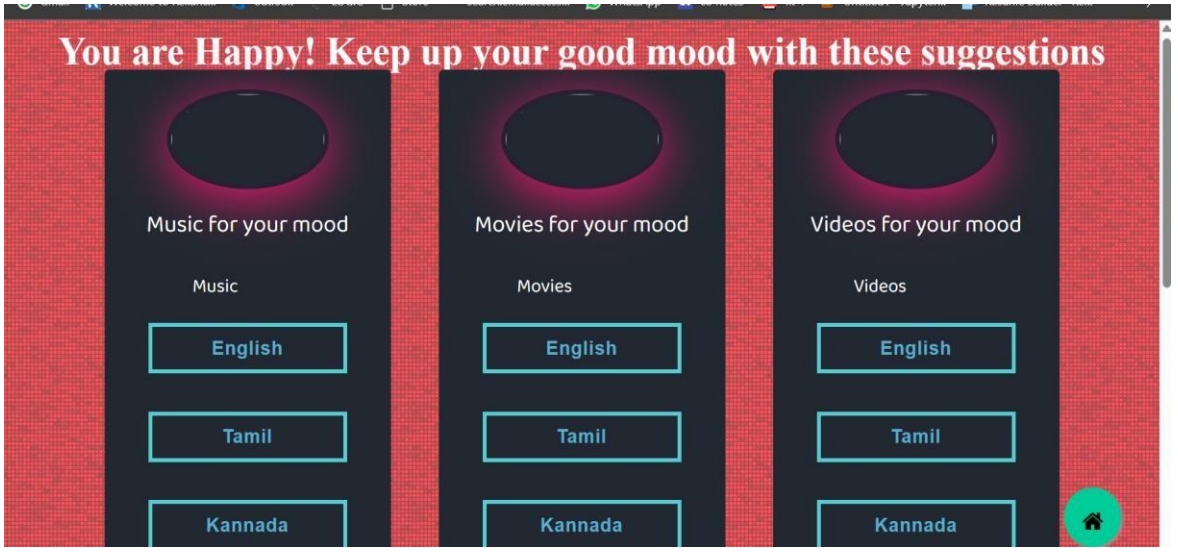


Fig 6.5: user is happy, so the page will be redirected to happy site .

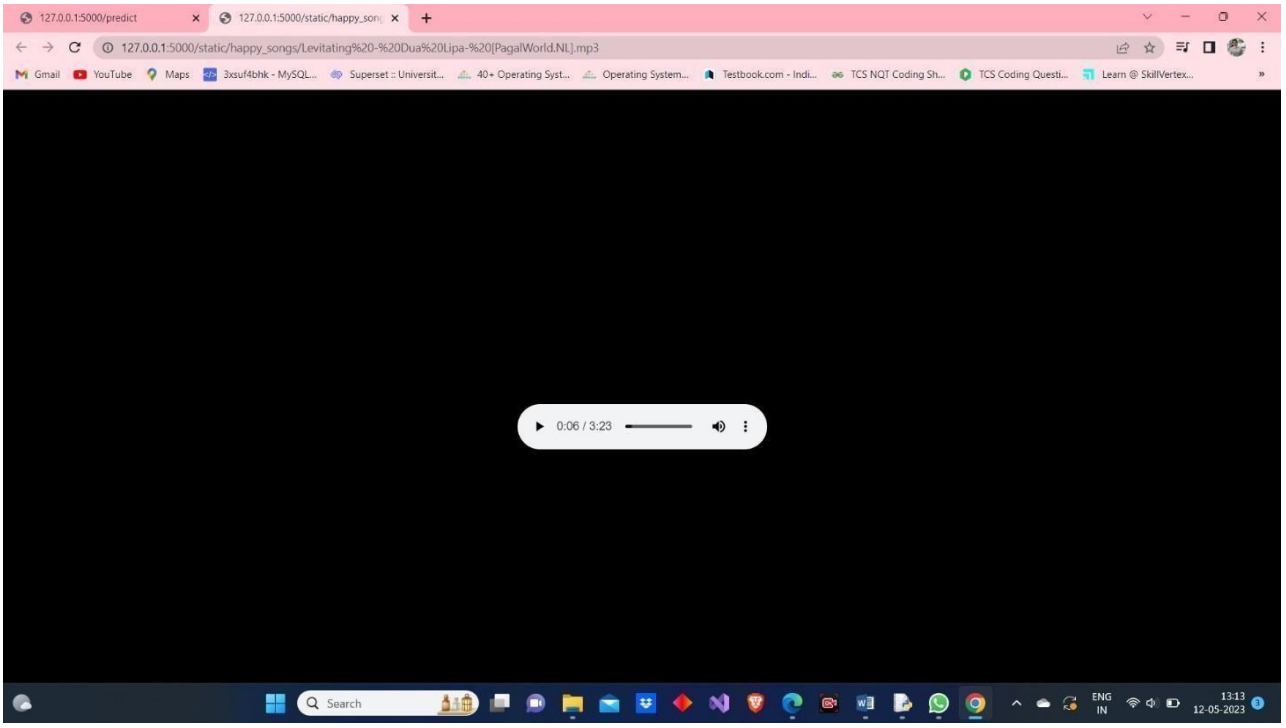


Fig 6.6: Happy songs are played.

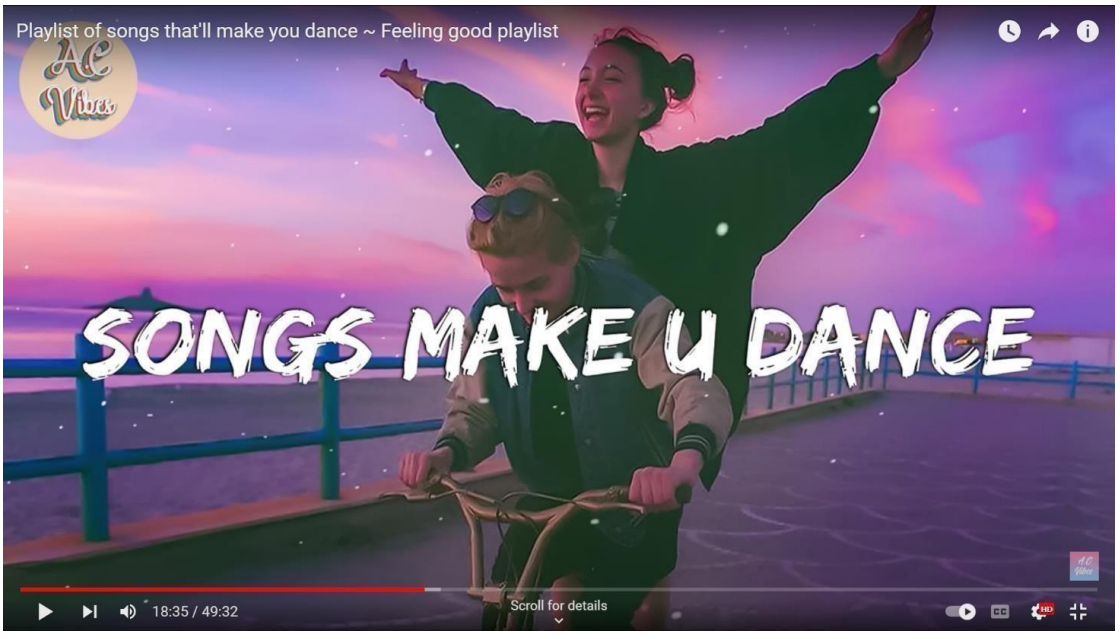


Fig 6.7: YouTube videos related to happy mood can be watched.

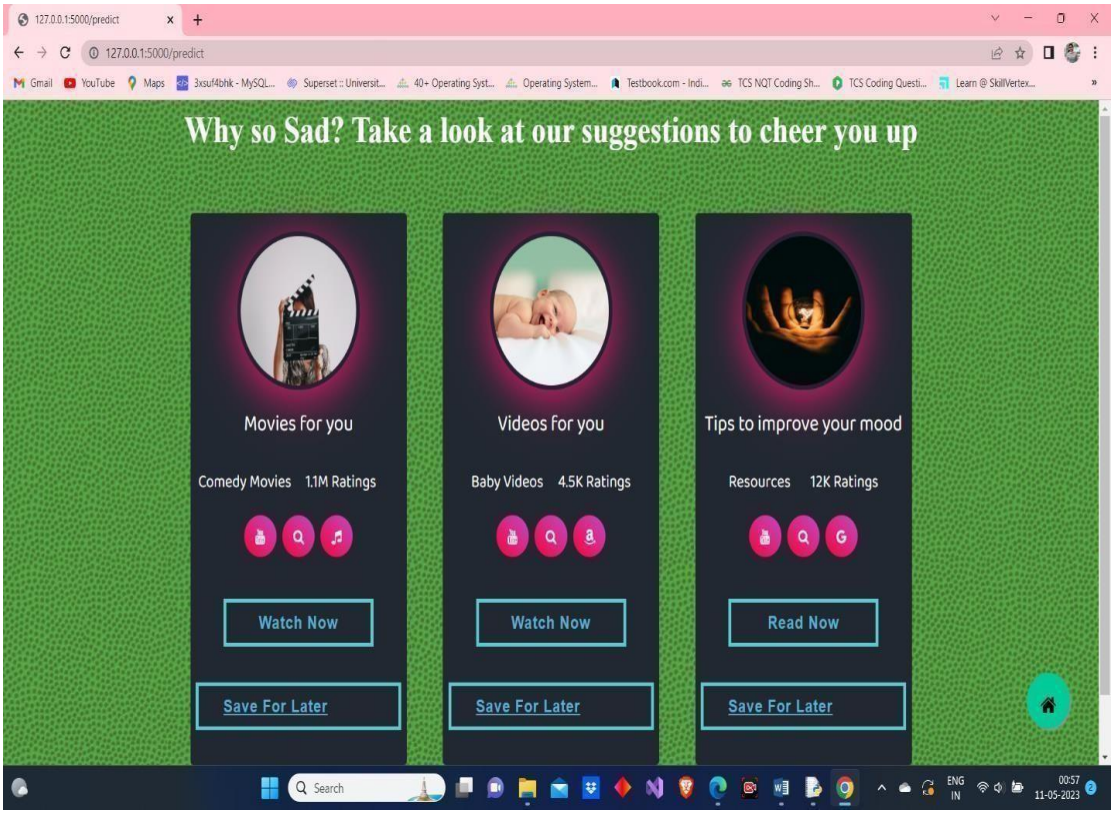


Fig 6.8: The page will be redirected to sad site as shown.

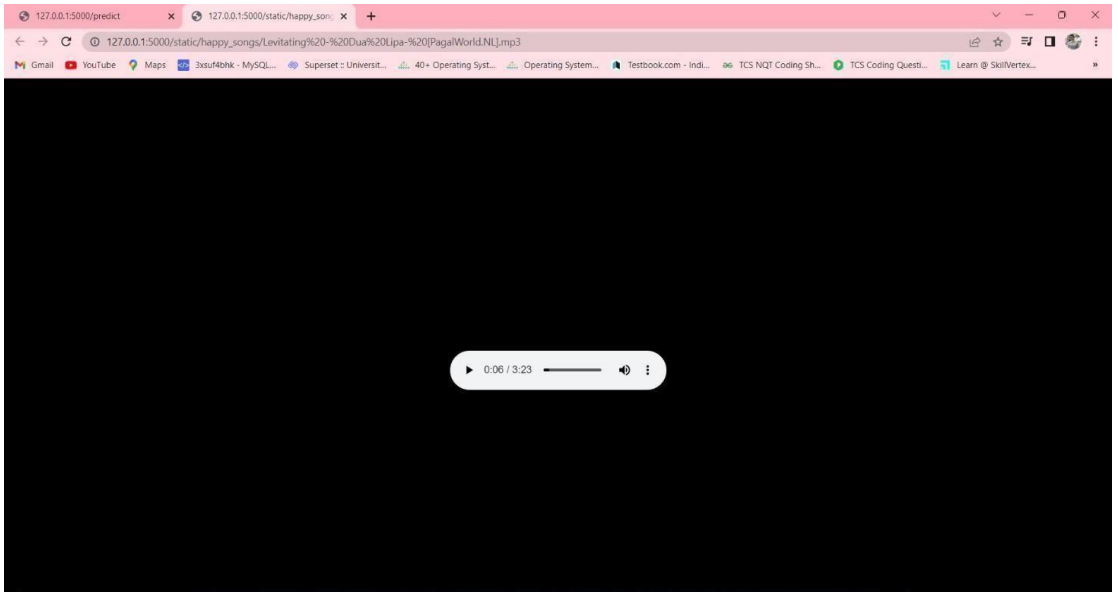


Fig 6.9: Calming songs are played.

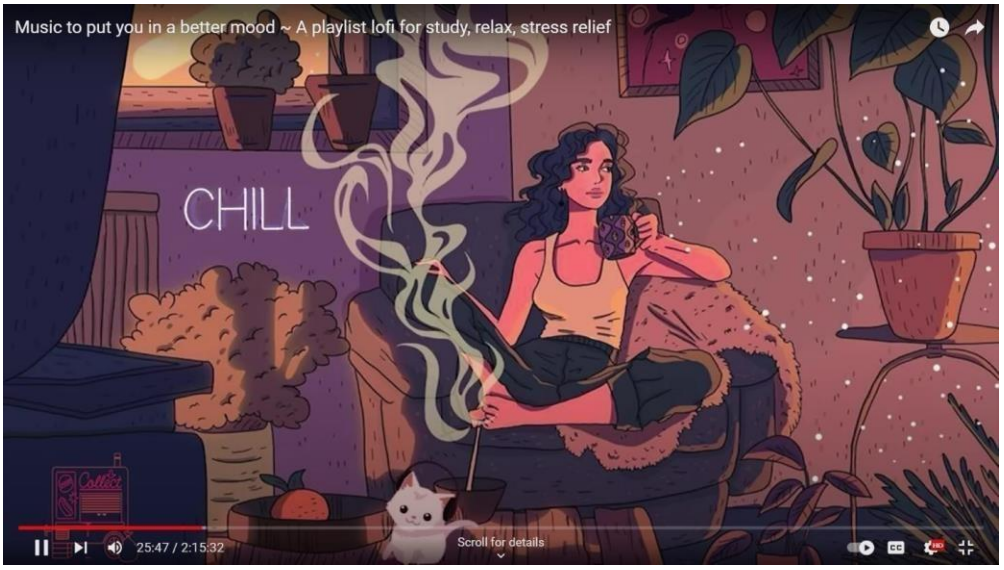


Fig 6.10: YouTube music to put the person in better mood.

Fig 6.11: Funny videos to watch.

6.5 Chat-Bot

The below figure shows another feature in our application called chat-bot. This can be used by the userto interact with application.

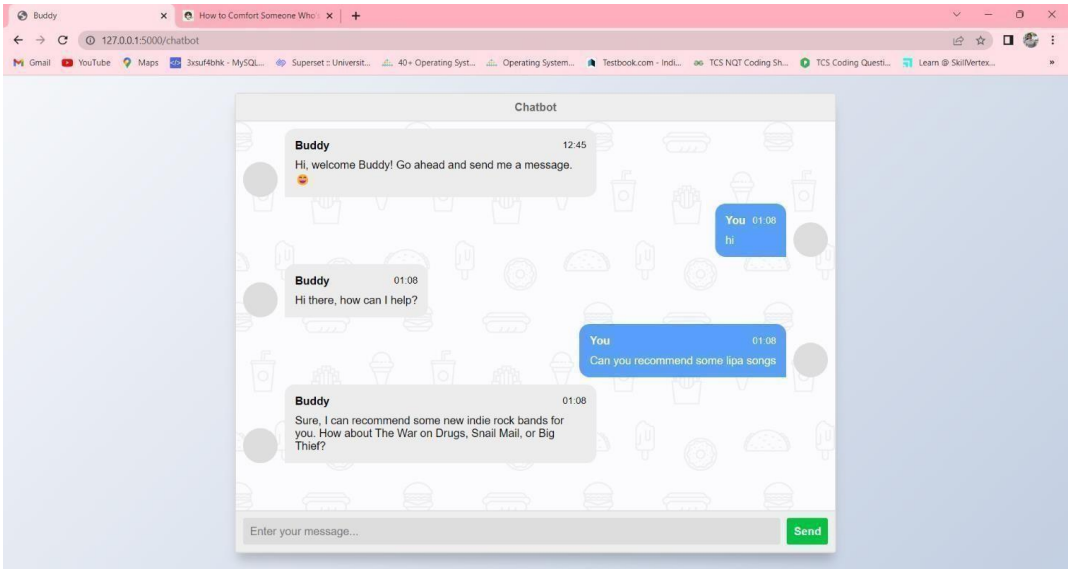


Fig 6.12: In Chabot, user can communicate. .

CONCLUSION

7.1 Conclusion

In this project, we presented a music recommendation system based on emotion detected. The system uses a two-layer convolution network model for facial emotion recognition. The model classifies 7 different facial emotions from the image dataset. The model has comparable training accuracy and validation accuracy which convey that the model is having the best fit and is generalized to the data. We also recognize the room for improvement. It would be interesting to analyze how the system performs when additional emotions are taken into consideration. User preferences can be collected to improve the overall system using collaborative filtering. We plan to address these issues in future work.

7.2 Scope of Future Enhancement

The music player based on facial recognition system is highly essential for all the person in modern day life ecology. This system is further enhanced with benefit able features for upgrading in future. The methodology of enhancement in the automatic play of songs is done by detection of the facial expression. The facial expression is detected by programming interface with the RPI camera. Our system currently excludes emotions such as disgust and fear. However, an alternative method could incorporate these emotions to enhance automatic music playback.

REFERENCES

[1] X. Zhang, G. Qi, X. Fu and N. Zhang, "Robust Emotion Recognition Across Diverse Scenes: A Deep Neural Network Approach Integrating Contextual Cues," in IEEE Access, vol. 11, pp. 73959-73970, 2023, doi: 10.1109/ACCESS.2023.3296316.

[2] A. R. Jafari, G. Li, P. Rajapaksha, R. Farahbakhsh and N. Crespi, "Fine-Grained Emotions Influence on Implicit Hate Speech Detection," in IEEE Access, vol. 11, pp. 105330-105343, 2023, doi: 10.1109/ACCESS.2023.3318863.

[3] A. M. Yildiz et al., "FF-BTP Model for Novel Sound-Based Community Emotion Detection," in IEEE Access, vol. 11, pp. 108705-108715, 2023, doi: 10.1109/ACCESS.2023.3318751.

[4] F. Alrowais et al., "Modified Earthworm Optimization With Deep Learning Assisted Emotion Recognition for Human Computer Interface," in IEEE Access, vol. 11, pp. 35089- 35096, 2023.

[5] M. Al-Hashedi, L. -K. Soon, H. -N. Goh, A. H. L. Lim and E. -G. Siew, "Cyberbullying Detection Based on Emotion," in IEEE Access, vol. 11, pp. 53907-53918, 2023, doi: 10.1109/ACCESS.2023.3280556.

[6] J. Heredia et al., "Adaptive Multimodal Emotion Detection Architecture for Social Robots," in IEEE Access, vol. 10, pp. 20727-20744, 2022, doi: 10.1109/ACCESS.2022.3149214.

[7] M. T. Teye, Y. M. Missah, E. Ahene and T. Frimpong, "Evaluation of Conversational Agents: Understanding Culture, Context and Environment in Emotion Detection," in IEEE Access, vol. 10, pp. 24976-24984, 2022, doi: 10.1109/ACCESS.2022.3153787.

[8] N. Aslam, F. Rustam, E. Lee, P. B. Washington and I. Ashraf, "Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model," in IEEE Access, vol. 10, pp. 39313-39324, 2022, doi: 10.1109/ACCESS.2022.3165621.

[9] C. -L. Hwang, Y. -C. Deng and S. -E. Pu, "Human–Robot Collaboration Using Sequential- Recurrent-Convolution-Network-Based Dynamic Face Emotion and Wireless Speech Command Recognitions," in IEEE Access, vol. 11, pp. 37269-37282, 2023, doi: 10.1109/ACCESS.2022.3228825.

[10] B. Li, H. Ren, X. Jiang, F. Miao, F. Feng and L. Jin, "SCEP—A New Image Dimensional Emotion Recognition Model Based on Spatial and Channel-Wise Attention Mechanisms," in IEEE Access, vol. 9, pp. 25278-25290, 2021, doi: 10.1109/ACCESS.2021.3057373.

[11] M. -H. Hoang, S. -H. Kim, H. -J. Yang and G. -S. Lee, "Context-Aware Emotion Recognition Based on Visual Relationship Detection," in IEEE Access, vol. 9, pp. 90465- 90474, 2021, doi: 10.1109/ACCESS.2021.3091169.

[12] F. Ren and T. She, "Utilizing External Knowledge to Enhance Semantics in Emotion Detection in Conversation," in IEEE Access, vol. 9, pp. 154947-154956, 2021, doi: 10.1109/ACCESS.2021.3128277.

[13] N. Samadiani, G. Huang, Y. Hu and X. Li, "Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features," in IEEE Access, vol. 9, pp. 35524- 35538, 2021, doi: 10.1109/ACCESS.2021.3061744.

[14] N. Aslam, F. Rustam, E. Lee, P. B. Washington and I. Ashraf, "Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model," in IEEE Access, vol. 10, pp. 39313-39324, 2022, doi: 10.1109/ACCESS.2022.3165621

[15] C. -L. Hwang, Y. -C. Deng and S. -E. Pu, "Human–Robot Collaboration Using Sequential- Recurrent-Convolution-Network-Based Dynamic Face Emotion and Wireless Speech Command Recognitions," in IEEE Access, vol. 11, pp. 37269-37282, 2023, doi: 10.1109/ACCESS.2022.3228825.