Multimodal Emotion Recognition

Aniruddha Patil[1],Anuradha Thakare[2], Sanjay Talbar[3], Namish Arora[4], Atherva Bhojane[5], Tejashree Chougule[6]
Department of Computer Science(AI & ML)
[1,2,3,4,5,6]Pimpri-Chinchwad College of Engineering, Pune
[1]aniruddha.patil22@pccoepune.org, [2]anuradha.thakare@pccoepune.org,
[3]sanjay.talbar@pccoepune.org, [4]namish.arora22@pccoepune.org,
[5]atherav.bhojane22@pccoepune.org, [6]tejashree.chougule22@pccoepune.org

1.ABSTRACT:

In today's busy world, people face many stress related problems coming from various aspects of life, such as demanding work schedules, competing for achievements, promotions and managing different responsibilities which are on their shoulders. Studies suggest that numerous individuals who have encountered major life challenges, may display elevated levels of negative feelings.These negative emotions may force any individual to take an action which might cause harm to them and their families like a succide attempt. Dealing with life after facing major stress can be really tough for people and their support networks. In response to this issue, we suggest the creation of a system , designed to identify and understand the emotions of the people who are affected by stress and negative emotions. The system will analyze various psychological parameters, like facial gestures, eye movements - blinking rate, pupil dilation, EEG signals, etc. These signals will provide valuable information about the current emotional state of an individual and will help in understanding them better.

2.KEYWORDS:

emotions, valence, arousal dimensions, EEG, EMG, ECG,k-nearest neighbors, support vector machines, Recurrent Neural Networks, Convolution Neural Networks

3.INTRODUCTION:

The emotions are discrete and fundamentally different constructs, emotions can be characterized on a dimensional basis in groupings. Humans' subjective experience is that emotions are clearly recognizable in ourselves and others.

Emotion recognition is the process of correctly identifying various emotions on the basis of various parameters. It is crucial in understanding human communication and daily interactions. It involves analyzing emotional response characteristics and signals to assess human emotional states. These emotional expressions are often quite ambiguous and occur rapidly (even without our awareness) and therefore depend upon one another and context to be accurately identified. Emotion recognition guides response and action toward potential friendly or threatening others. In order to identify emotions, we process both static and dynamic cues, such as facial expressions and bodily gestures. Emotions can be represented using two models: the discrete model, which categorizes emotions into primary states such as anger, disgust, fear, happiness, sadness, and surprise; and the dimensional model, which maps emotions onto valence and arousal dimensions.

Emotion recognition research involves capturing both physiological and non-physiological reactions to emotional states. Physiological reactions include skin temperature, respiratory rate, and heart rate, while non-physiological reactions encompass posture and facial expressions,eye movements. Researchers typically employ various sensors and devices to collect these data signals, such as EEG, Electrocardiogram (ECG), and Electromyogram (EMG). Multimodal emotion recognition integrates multiple data sources and features to enhance recognition accuracy.

According to theories of "basic emotions," like anger, sadness, or fear, are activated when the brain evaluates a stimulus or event in relation to the individual's goals or survival instincts. These emotions are believed to have unique biological functions, expressions, and significance, separate from each other.

For the past several years, systems using artificial intelligence have been “learning” to detect and distinguish human emotion by associating feelings such as anger, happiness, and fear, with facial and bodily movements, words, and tone of voice.

The accuracy of emotion recognition is usually improved when it combines the analysis of human expressions from multimodal forms such as texts, physiology, audio, or video. Different emotion types are detected through the integration of information from facial expressions, body movement,eye movements, gestures and speech.

Emotion recognition systems often benefit from considering contextual information and situational cues. Understanding the context in which emotions occur can improve the accuracy of emotion recognition. For example, interpreting facial expressions in the context of a conversation or analyzing speech prosody along with the content of spoken words.

Various machine learning algorithms are employed to analyze extracted features and classify them into different emotional states. These algorithms include traditional methods such as Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbors (k-NN), as well as more advanced techniques like deep learning.
Deep learning approaches, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have gained popularity in recent years for their ability to automatically learn hierarchical representations from raw data.

Emotion recognition technology is being integrated into various systems and applications to make them emotion-aware. For example, emotion-aware virtual assistants can adapt their responses based on the user's emotional state, while emotion-aware educational software can provide personalized learning experiences tailored to the student's emotions.
As emotion recognition technology becomes more pervasive, there is growing awareness of the ethical implications associated with its use. Concerns include privacy issues related to the collection and analysis of sensitive emotional data, potential biases in the algorithms used for emotion recognition, and the impact of emotion-aware systems on user autonomy and well-being.

4.RELATED RESEARCH PAPER:

Sr.No	Paper Title, Authors and Publication Year	Algorithm/Technology	Strength	Limitations
1	Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets U. Bilotti , C. Bisogni , M. De Marsico , S. Tramonte IEEE, April 2023	CNN with multiple inputs from the dataset.	Good accuracy in classification based on video inputs (about 93%).	Lack of a structure capable of channeling a greater number of data modalities. No proper addressing of evaluating the performance of trained models across different datasets.
2	Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning Hoai-Duy Le; Guee-Sang Lee; Soo-Hyung Kim; Seungwon Kim; Hyung-Jeong Yang IEEE, 13 February 2023	Transformer-based fusion	Higher accuracy in correctly classifying raw videos	More time consuming as there is no filter for removing the duplicated frames for reducing computational cost
3	Multimodal Emotion Recognition in Response to Videos Mohammad Soleymani; Maja Pantic; Thierry Pun IEEE 06 December 2011	Modality fusion strategy and support vector machine	Model which did not require direct user inputs to classify different emotions based on pupillary reflex and EEG.	The results were based on a fairly small dataset.

4	A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition Tao Meng , Yuntao Shou , Wei Ai , Jiayi Du , Haiyan Liu , Keqin Li Elsevier 07 September 2018	GNN(Graphical Neural Networks)	Good accuracy as compared to previous study Improvement in the generalization ability of the model	Limitation of dataset,only 2 datasets used- IEMOCAP, ME-LD
5	Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition Wei Liu; Jie-Lin Qiu; Wei-Long Zheng; Bao-Liang Lu IEEE, 05th April, 2021	Weighted sum fusion Attention Based fusion	Improved accuracy on SEED-V dataset, The model had more homogenous features.	The CCA metric used, can fuse only two modalities,which may not be suitable for real life situations
6	Deep Multimodal Emotion Recognition on Human Speech: A Review Panagiotis Koromilas Theodoros Giannakopoulos Applied Science, 28th August 2021	Random sub sampling	Model is robust against different kinds of outliers in any type of speech input.	Model is engineered on specific dataset, Not trained on cross-dataset

7	Multimodal emotion recognition from expressive faces, body gestures and speech George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis & Kostas Karpouzis AIAI 2007 Conference paper	Bayesian Classifier, For emotion classification of different emotions	Increased recognition rates by fusing multimodal data.	Classification was done for eight different emotions which might not be suitable for classifying emotions from different expressivity of the same gesture.
8	Multi-modal emotion recognition using EEG and speech signals Qian Wang, Mou Wang, Yan Yang & Xiaolei Zhang 29th October 2022	Feature fusion, decision level fusion and Multilayer Perceptron Network (MLP)	Good accuracy in correctly classifying the emotions.	The model does not work in real life. Different scenarios like the model gets confused while classifying sad and neutral emotion and signals are sensitive to noise.
9	A systematic survey on multimodal emotion recognition using learning algorithms Naveed Ahme,Zaher Al Aghbari, Shini Girija February 2022	SVM, Feature level fusion, Decision level fusion, Hybrid level fusion	Accuracy was good (97%) from feature level fusion technique ,on multimodal features- EEG and facial expressions	No generalized classifier for predicting various emotions of different people,

10	Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions Geetha A.V., Mala T., Priyanka D.,Uma E. August 2022	Deep learning algorithms like fusion techniques	Model was able to work on various modalities like EEG signals ,eye reflex, speech and video inputs	Emotional misclassification Representing complex emotions like pain,sorrow was difficult
11	MULTIMODAL EMOTION RECOGNITION Nicu Sebe, Ira Cohen, and Thomas S. Huang September 2011	KNN, HMM, neural network,bayesian network	A combination of low-level features, high-level reasoning, and natural language processing is likely to provide the best emotion inference.	Invasion of human privacy which is against the social ethics.
12	Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning Samarth ,Sarthak Tripathi, Homayoon Beigi	3D-CNN , text-CNN and openSMILE for audio feature extraction	Faster training and inference time of the model	Model is trained on only one dataset(IEMOCAP)
13	A Multimodal Emotion Recognition Model integrating speech, video and MoCAP Ning Jia, Chunjun Zheng & Wei Sun 13 April 2022	Facial Motion Speech Emotion Recognition (FM-SER),	higher recognition accuracies in multimodal features.	Low accuracy in single modality, dataset is very huge, difficult to process

14	Multimodal Emotion Recognition using Deep Learning Sharmeen M.Saleem Abdullah Abdullah, Siddeeq Y. Ameen Ameen, Mohammed A. M. Sadeeq, Subhi Zeebaree May 2021	Deep belief network(DBn), CNN,RNN	this method generates different accuracy on different posed databases, to approximately 63% when dealing with DEAP data set while with DECAF data set the accuracy is less about 57%	Emotional variations can be observed in physiological signals for a very short period of around 3-15 seconds.
15	Facial emotion recognition using deep learning: review and insights, Procedia Computer Science, Wafa Mellouk, Wahida Handouzi 2020	CNN, CNN-LSTM, VGGNet network	Higher accuracy achieved on CK+, JAFFE, BU-3DFE datasets.	More complex emotions are difficult to identify
16	A review of multimodal emotion recognition from datasets Bei Pan, Kaoru Hirota, Zhiyang Jia, Yaping Dai	ML - SVM,RF DL-CNN, graphical neural Networks(GNN)	High accuracy on SVM and fusion level techniques	Misclassification of emotions are there
17	Multimodal Emotion Recognition From EEG Signals and Facial Expressions Shuai Wang; Jingzi Qu; Yong Zhang; Yidie Zhang 31 March 2023	CNN	proposed model can effectively recognize emotions Accuracy- valence dimension classification is 96.63% & arousal dimension classification is 97.15%	Can introduce more modalities to enhance the model.

18	Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions Geetha A.V. , Mala T., Priyanka D. , Uma E. 21 December 2023	Deep Learning	review thoroughly investigates the key elements that constitute its pipeline, including foundational emotion theories, data preprocessing, feature extraction, feature engineering techniques	Heterogeneity of modalities: This heterogeneity arises from the diverse nature of information conveyed through different modalities, including varying data types, distinct feature representations, and temporal dynamics.
19	Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network Ngoc-HuynhHo,Hyung-Jeong Yang; Soo-Hyung Kim; Gueesang Lee 30 March 2020	RNN (recurrent neural network) & Self-to-MultiHead attention mechanism	the combination of the two modalities achieves better performance than using single models	There are dynamic-size inputs used rather than the whole this can produce very largely non-informative samples if there is a big gap of temporal length between trials
20	Multimodal Emotion Recognition using Deep Learning Architectures Hiranmayi Ranganathan, Shayok Chakraborty and Sethuraman Panchanathan 2016	CDBN,DBN	DemoDBN models perform better than the state of the art methods for emotion recognition using popular emotion corpora	Data from one or more modalities may be absent, to develop models that will continue to perform and successfully recognize emotions even when one or more modalities are absent

5.COMBINED RELATED WORK:

- **Strength:**
  1. **Accurate Across Different Inputs:**
     a. The model is really good at figuring out emotions from videos, getting it right about 93% of the time.
  2. **Works Well with Raw Videos**:

    a. It's even better at understanding emotions in videos that haven't been processed beforehand.

3. **Automatic Emotion Recognition:**
    a. It can figure out how someone's feeling just by looking at their pupils and EEG signals, without needing direct input from the user.
4. **Better Adaptation to Different Situations:**
    a. This model is getting better at understanding emotions in a wide range of situations and datasets compared to older versions.
5. **Consistent Results with Clear Features:**
    a. It's really good at picking up patterns in data, especially in the SEED-V dataset, where it performs consistently well.
6. **Handles Different Speech Patterns Well**:
    a. It's robust enough to understand emotions in speech even when there are unusual patterns or outliers.
7. **Combining Different Data Boosts Accuracy:**
    a. By putting together information from different sources, like EEG and facial expressions, it gets better at recognizing emotions.
8. **Gets Emotions Right:**
    a. It's pretty accurate at recognizing how people are feeling, which is important for understanding human behavior.
9. **Combining Data Features Effectively:**
    a. It does a great job of putting together different kinds of information, like EEG signals and facial expressions, to get really accurate results.
10. **Works with Different Types of Data:**
    a. It can handle all sorts of data, from brain signals to video clips, making it versatile for different applications.

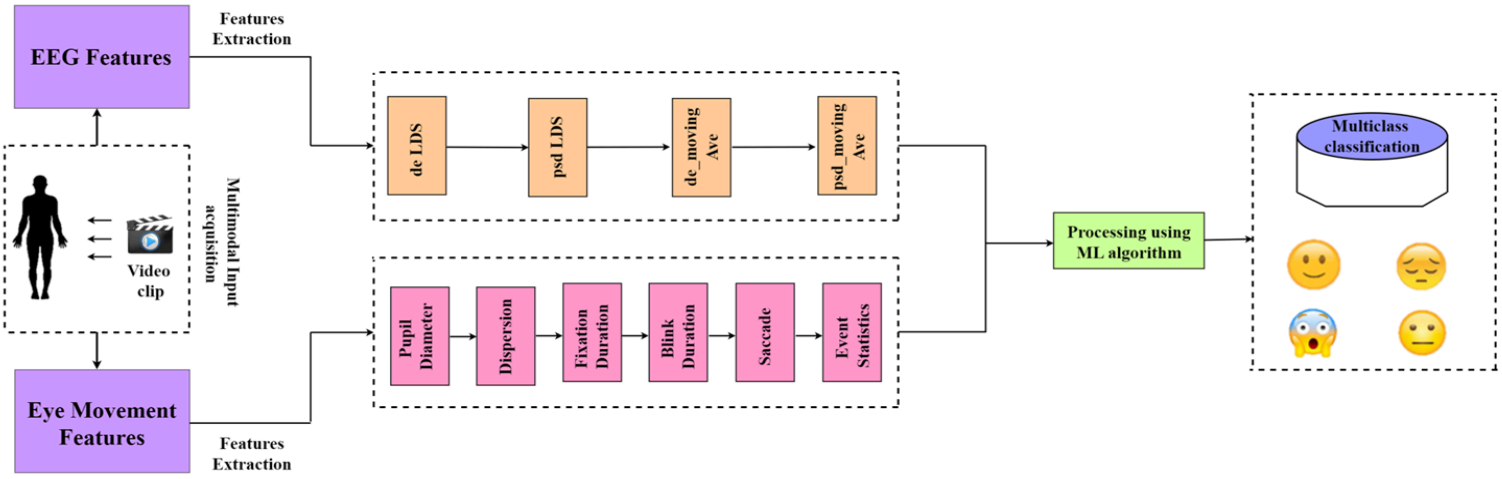
- ## Limitations:

1. **Limited Data Modalities Handling**:
    a. The model lacks a structure to handle multiple types of data sources effectively.
2. **Inadequate Evaluation Across Diverse Datasets**:
    a. There's a need for better methods to evaluate how well the model performs across different datasets.
3. **Time-Consuming Due to Duplicate Frames**:
    a. It takes more time to process because it doesn't filter out duplicated frames, which could reduce computational load.
4. **Small and Limited Datasets**:

    a. The results are based on a small dataset, only using two datasets (IEMOCAP and ME-LD), which limits the model's understanding of emotions.

5. **Limited Modality Fusion**:
    a. The fusion method used can only combine two types of data, which might not be practical for real-world situations.
6. **Engineered on Specific Data**:
    a. The model is designed based on a specific dataset, potentially limiting its performance in broader contexts.
7. **Lack of Cross-Dataset Training**:
    a. It's not trained on data from different sources, which could affect its ability to generalize to new situations.
8. **Limited Emotion Classification**:
    a. The model is trained to classify only eight specific emotions, which might not cover the full range of human emotional expressions.
9. **Real-Life Performance Issues**:
    a. The model struggles in real-life scenarios, such as confusing sad and neutral emotions, being sensitive to noise, and lacking a generalized approach for predicting emotions across different individuals.
10. **Difficulty Representing Complex Emotions**:
    a. Complex emotions like pain and sorrow are challenging for the model to understand and classify accurately.

6.ARCHITECTURE DIAGRAM:



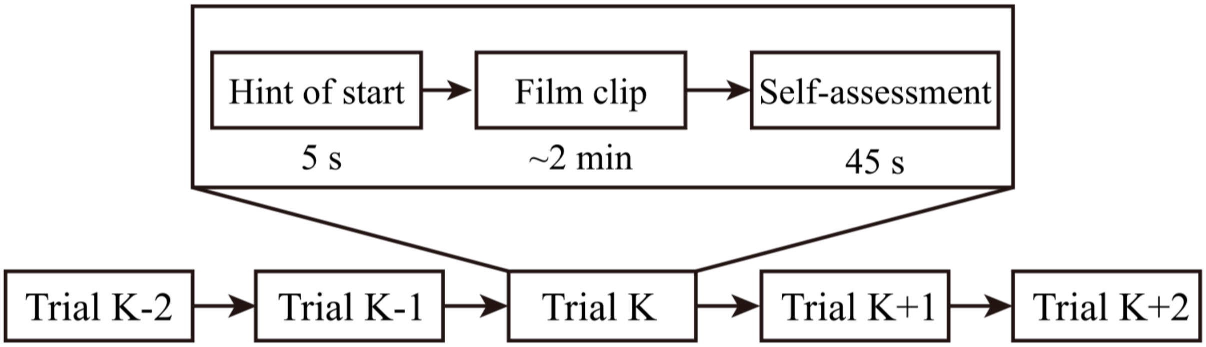
Architecture Diagram

7.DATASET DESCRIPTION:

Dataset chosen to work on Multimodal Emotion Recognition is the SEED IV Dataset. The given dataset is collected for various purposes using EEG signal and Eye tracker. By doing experiments on participants using EEG cap and Eye tracker, data collected. Data has been transformed to 4 files namely, feature1.csv, feature2.csv, feature.csv, dataset.csv. Dataset.csv contains all these 3 features in combined format. The feature.csv contains data about different sessions conducted on different days for participants.

7.1.Experimental setup

Seventy-two film clips were meticulously curated through a preliminary investigation, each wielding the power to evoke a spectrum of emotions: Happy, Sad, Fear, and neutrality. Fifteen individuals were enlisted as participants in the study. Across the course of the experiment, each participant engaged in three distinct sessions, conducted on separate days. Within each session, a series of 24 trials was administered, yielding a comprehensive dataset for analysis.During each trial, participants viewed a selected film clip while their EEG signals and eye movements were simultaneously recorded. This data collection was facilitated through the utilization of the 62-channel ESI NeuroScan System in conjunction with SMI eye-tracking glasses. Trial Diagram is as follows:



7.2.Feature Extraction

The raw EEG data undergo initial downsampling to a 200 Hz rate. Subsequent noise reduction and artifact removal are achieved through bandpass filtering within the 1 Hz to 75 Hz range. Features such as power spectral density (PSD) and differential entropy (DE) are then extracted across five frequency bands: delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz), and gamma (31-50 Hz). PSD and De can be calculated as follows:

$$\textbf{PSD} = E[x^2]$$

$$\textbf{DE} = -\int_{-\infty}^{\infty} P(x)\ln(P(x))dx \quad \textbf{DE} = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp\frac{(x-\mu)^2}{2\sigma^2} \ln(\frac{1}{\sqrt{2\pi\sigma}} \exp\frac{(x-\mu)^2}{2\sigma^2})dx = \frac{1}{2}\ln 2\pi e\sigma^2$$

Eye movement extracted through SMI eye-tracking glasses contain various parameters. There are as follows:

Pupil Diameter, Dispersion, Fixation Duration, Blink Duration, Saccade, Event Statistics.

7.3.Dataset Summary

Dataset contains 52 columns and 1079 rows. 15 participants participated in 3 different sessions conducted on 3 different days. Each participant goes through 24 trials.

SR. NO.	EEG signals and eye movement parameters	Rows and columns
1	1 sessions	1 - 360 rows
2	2 session	361 - 720 rows
3	3 session	721 - 1079 rows
4	Pupil Diameter	1 - 12 columns
5	Dispersion	13 - 16 columns
6	Fixation Dispersion	17 - 18 columns
7	Saccade	19 - 22 columns
8	Event Statics	23 - 31 columns
9	DE_LDS	32 - 36 columns
10	PSD_LDS	37 - 41 columns
11	DE_MovAv	42 - 46 columns
12	PSD_MovAv	47 - 50 columns
13	Classified Emotion	51 column

Here, PSD AND DE are the features of EEG signals. Linear dynamic systems (LDS) and moving averages are two different approaches to filter out noise and artifacts that are unrelated to EEG features which can be used for smoothing purpose.

8.DISCUSSION ON RESULTS:

The results obtained as per the various machine learning applied on the dataset. We implemented various algorithms successfully which gives accurate and precise output. To be more accurate we need to enhance and add more inputs in the dataset. Currently, random forest and logistic regression gives

Sr.no	ML Algorithm used	Parameters	Maximum accuracy achieved
1	**Decision Tree**	Criteria : Entropy random state : 50 and 100 min_sample_leafs : 10	56.2 %
2	**Random Forest**	n_estimators : 25	67.03 %
3	**SVM**	–	–
4	**Naive bayes**	Criteria: ratio = 0.9 90% training, 10% testing	26.74 %

5	**KNN**	PCA(n_components = 2)
		Random_state = 45
		Test_size = 0.2
		n_neighbours = 5
			34.26 %

6	**Logistic Regression**	test_size=0.5,
		random_state=20
			67.59 %

9.CONCLUSION:

This paper presents a novel approach to multimodal emotion recognition by integrating EEG signals and eye movements. This paper extracts features from both modalities, enabling a comprehensive understanding of emotional states. Concurrently, ML classification algorithms are employed to capture spatial features from EEG signals. These ML algorithms help to classify the 4 emotions (happy, neutral, angry, sad). Experimental findings demonstrate the accuracy of different ML algorithms, showcasing performance in multimodal emotion recognition. Future endeavors will focus on optimizing pre-training models for facial expression feature extraction, streamlining model operation, and exploring additional modalities, such as non-physiological signals, to further enhance the multimodal emotion recognition framework.

10.REFERENCES:

1.Emotion Classification https://en.wikipedia.org/wiki/Emotion_classification

2.Multimodal Emotion Recognition From EEG Signals and Facial Expressions:
https://ieeexplore.ieee.org/abstract/document/10089483

3.Multimodal Emotion Recognition in Response to Videos
https://ieeexplore.ieee.org/abstract/document/6095505

4.A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition
https://www.sciencedirect.com/science/article/abs/pii/S0925231223012328

5.Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition
https://ieeexplore.ieee.org/abstract/document/9395500

6.Deep Multimodal Emotion Recognition on Human Speech: A Review
https://www.mdpi.com/2076-3417/11/17/7962

7.Multimodal emotion recognition from expressive faces, body gestures and speech
https://link.springer.com/chapter/10.1007/978-0-387-74161-1_41

8.Multimodal emotion recognition using EEG and speech signals
https://www.sciencedirect.com/science/article/pii/S0010482522006503

9.Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions
https://www.sciencedirect.com/science/article/abs/pii/S1566253523005341

10.https://link.springer.com/article/10.1007/s11042-022-13091-9

11.https://www.sciencedirect.com/science/article/pii/S1877050920318019

12.https://www.sciencedirect.com/science/article/abs/pii/S092523122300989X

13.https://ieeexplore.ieee.org/abstract/document/10089483

14.https://www.sciencedirect.com/science/article/pii/S1566253523005341#sec10