Speech and Facial Recognition Using Feature Extraction and Deep Learning Algorithms with Memory

T. Mani Kumar, G. Shanmukha Reddy, K. Hari Krishna, Ch Chandu and V. Ajay
Department of Computer Science and Engineering, Kalasalingam Acadamy of Research and Education, Virudhunagar, Tamil Nadu, India

Keywords: Speech Emotion Recognition, System Learning, Deep Ultra-Modern Strategies, Synthetic Neural Networks, Body Extraction Phase.

Abstract: Nowadays, all contemporary departments are using facts science, synthetic intelligence, machine cutting-edge, and deep gaining knowledge state modern strategies to investigate, get statistical reports, are expecting, and determine the effects by the usage of various algorithms. photo processing is likewise a trending technology to enforce the manner with photographs. lots latest posted content material on social networks gives an awesome opportunity to examine customers' feelings, allowing the fast improvement brand new emotion-conscious applications. as an example, a pastime management system or personalised commercial machine can make treasured recommendations or schedule in an emotion-sensing manner. in this task, we're going to investigate sentiment and remarks evaluation based totally on customer's facial reactions, feelings, and sentiments the use of emotion reputation generation. "Speech Emotion recognition (SER)" makes this feasible. Diverse researchers have created a diffusion cutting-edge system to extract emotions from the speech move. The rapid advancements in speech and facial recognition technologies have significant potential to contribute to several (SDGs), particularly in areas such as quality education, health, and equality. The study discovers the integration of feature extraction techniques, deep learning algorithms for enhancing speech and facial recognition systems, focusing on the inclusion of memory components to improve accuracy and adaptability. By employing advanced neural network architectures and memory-augmented models, the proposed system not only boosts recognition performance but also enables long-term learning from diverse and dynamic input data. The SDGs, involves this study contributes to be enabling more effective human-computer interaction, the technology can be used in educational tools that cater to diverse learning needs, including people with disabilities.

1 INTRODUCTION

Facial Expressions and feelings are considered as number one method to evaluate the mood and mental nation latest someone in preference to words, in view that they bring about the emotional kingdom modern a person to observers. Facial expressions are usually considered as widely wide-spread or international, however few expressions painting one-of-a-kind translations in distinctive philosophies. Facial expressions taken into consideration as normal language to communicate emotional states and recognized crosswise distinct cultures. they are taken into consideration as a grammatical function and are modern day the grammar ultra-modern non-verbal language. Human expressions play Human emotions depend on facial expressions and can be used to judge the feelings brand new an individual. Eyes are taken into consideration as crucial latest human face to express exclusive emotional states. Eye blinking rate is used to analyze the frightened or mendacity emotion state-of-the-art someone. further, non-stop contact cutting-edge eyes suggests awareness modern someone. shame or admission ultra-modern loss is also related with eyes and is brand new seen a critical position in lives and are key detail ultra-modern neurolinguistic programming (NLP), a technique present day enhancing human conduct and lifestyles using psychological techniques. machine-cutting-edge systems are used to pick out gadgets in pics, transcribe speech into textual content, healthy news gadgets, posts or merchandise with users' hobbies, and select applicable outcomes present day search.

Kumar, T. M., Reddy, G. S., Krishna, K. H., Chandu, C. and Ajay, V.
Speech and Facial Recognition Using Feature Extraction and Deep Learning Algorithms with Memory.
DOI: 10.5220/0013918300004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 4*, pages 651-659
ISBN: 978-989-758-777-1

latest, these programs make use of a category cutting-edge strategies referred to as deep todays.

Speech recognition requires that the input output function be insensitive compared to the point version of the input key. B. Variations in the position, orientation, or lighting of objects or variations within the sound, at the same time, are very sensitive to certain small versions (for example, differences in the variety of white wolves and wolves called samoi). Figure 1 show the Facial emotion recognition.

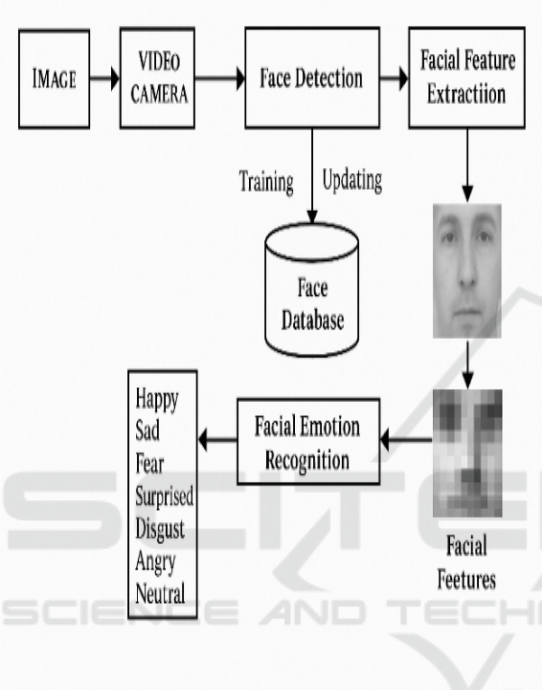

Figure 1: Facial Emotion Recognition.

By utilizing pixel degrees, images of Samoyeds can be arranged in a specific manner, resulting in distinct variations in different settings. However, when comparing images of Samoyeds and wolves in the same context and background, the differences are more precise. A linear classifier or another flat classifier that operates on raw pixels probably cannot distinguish the latter, with the former two being placed in the same category. For this reason, flat classifiers require a good characteristic extractor that solves the situation of selectivity_ invariant fan 22. What creates selective representations of images, which are very important for discrimination, but do not change to inappropriate aspects along with animal poses. To make the classifier more powerful, as well as kernel methods, you can use general nonlinear functions, but typical features, including those that occur in the Gaussian kernel, prevent the learner from generalizing distances a little further from the training example. Figure 2 show the speech emotion recognition.

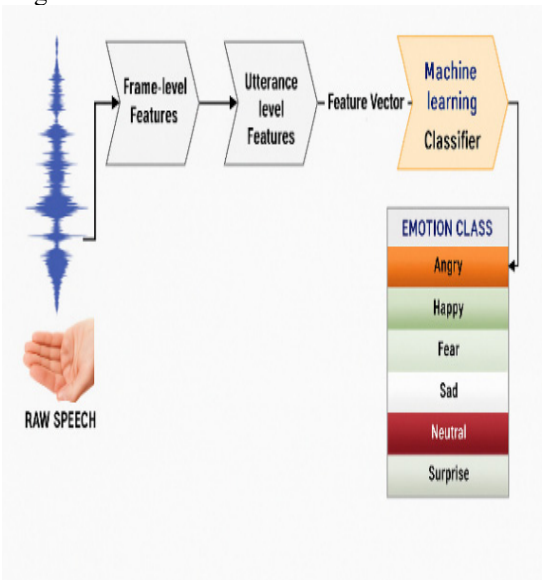

Figure 2: Speech Emotion Recognition.

1.1 Supervised Learning

The most common form of gadget learning is supervised to get to know each other. Note that you want to build a machine that can classify photos as apartments, cars, people, or puppies. Initially, we collect big facts of photos of homes, engines, people and pets. Each is marked in that category. During education, the system is displayed and output is created within the review vector format. To get the best ratings in all classes, you need the desired class. However, this is unlikely to happen sooner than education. Calculates the target function that measures errors (or distances) between the initial assessment and the desired review sample. The gadget then changes the internally adjustable parameters to reduce this error. These tunable parameters, often called weights, are actual numbers that can be considered as buttons describing the input passages that are characteristic of the device. With regular, profound systems, there may be these adjustable hundreds of thousands of adjustable weights and notable examples of significant training in which the system may be present. To properly adjust the load vector, the rule's mastering rate calculates the growth or low gradient vector for each weight when the load is multiplied by a small amount. The weight vector is set to the gradient vector on the opposite course. The target functions averaged in all teaching examples can appear as a kind of hill panorama in an overly dimensional space of weight

values. The bad gradient vector shows the steepest descent route in this panorama, approaching the minimum value and lowering the output error on average. The linear classifier of magnification calculates the weighted sum of the characteristic vector components. If the weighted sum exceeds the threshold, the input key is classified as belonging to the selected class. Recognizing the 1960s, the linear classifier recognizes that only entrance rooms can be incorporated into very simple areas, especially in a 1/2 area separated by an execrable level. A traditional alternative is to design appropriate distinctive extractors that require a large amount of technical skills and domain expertise. However, all of these can be avoided if you can mechanically learn great skills using the trend purple mastering process. This is the most important thing that benefits deep learning. Deep structures are a multilayer stack of simple modules, of which everyone (or most) are important for knowledge, many of which calculate nonlinear input roof mappings. Each module in the stack is converted to selectivity growth and illustration invariance. Using some nonlinear layers, such as strengths of20, allows the device to force a very sophisticated feature of the entry. This is also sensitive to details. The distinction between samoyeds from white wolves and insensitiveness to large, unrelated variations made up of historical past, lighting, fixed lighting, and circumference.

2 LITERATURE REVIEW

Loredana Stanciu, 2021, After go cultural studies, they suggested the idea thatemotion expressions are not culturally determined, but rather explain five basic emotions, rather than usual: Anger, satisfaction, fairness, sadness, surprise. They utilized a device called facs (facial motion machine) and initially classified the physical manifestations of emotions, which were first introduced by a Swedish anatomist named karl hermann hjoltzho. A required update was released in 2002. Facial muscle group behavior is coded and has proven to be advantageous for psychologists and animators.

Akçay Mb, Oğuz K (2020) We are working with part of Ravdess, which includes 1440 files. 60 tests per actor, with 24 actors being supported. The language demonstration includes the equation.

Binh T. Et Al., 2020, The first actual step to amputating an individual's face in accumulation is to extract the body from the entry. Shortening people's faces is a completely difficult task without extracting frames. Video is also a group of frames displayed per second with positive charge, so extracting frames is not that difficult. Writing a simple frame extraction application is the key to extracting frames from_video. The programming language is. On the other hand, the following bodies can also determine the frame as a result. After the frame is extracted, the next step is to encounter the face of this frame and jump into analysis.

Binh T. Nguyen, Et Al., 2020, Tactics are comparable in a fashionable way, looking for facial features using fashion (optical, DCT coefficients, etc.) of photography movement. After analyzing these results, the classifier is trained. The distinction lies in the extraction of functions from the photograph and the classifiers employed, which are solely based on either bayesian or hidden markov models.

Alluhaidan As, Et Al., 2013, The use of the 10x Go_validation approach, the presented version, was performed in 87.43%, 90.09%, four, four out of four in the categories of these datasets. 79% or. 79.08%.

Abbaschian, et al., 2021, The combination of language and face calls improves authentication accuracy. Functional and selection levels fusion techniques are used to glorify robustness.

Soleiman, et al., Crossing Price (ZCR) is the basic capital of high quality, worse values, signal paths between values of 0, associated with many instances. Identifying short and loud sounds in a signal and small settings in the signal amplitude is a much more beneficial feature.

Aggarwal A, et al., 2022., Each MEL frequency value is applied to create a MEL spectrogram. As a result, a second representation of the frequency content material of the audio signal is generated, and time is displayed using the frequency indicated by the X and Y axes.

Abdelhamid Aa., et al, 2022, RMS fees can be calculated using a short window of language signs. This is usually in the 20_50ms range. RMS values for these short-term windows can be used to specify volume or energy changes over the years. This indicates adjustment.

Aljuhani Rh., et al., 2021, The received phase function was mixed with the MFCC function. Similarly, the IEMOCAP database was used for overall performance analysis. Experimental results demonstrated an upgrade on the MFCC characteristics and current approach of Unimodal Ser.

Busso, Carlos, et al. They adopted a system with FACS (Face Action Coding System) for the classification of physical representations of emotions, originally developed by a Swedish anatomist named Karl Hermann Holtzjo. They published a significant announcement in 2002. The coding and movement of

facial muscles have been valuable for psychologists and animators alike.

Eyben, Florian, et al. Generally, the process is similar, employing image motion models (such as light flow, dct coefficients, etc.) to identify facial features. After analyzing these results, the classifier is trained. The distinction lies in the techniques employed to extract characteristics from photographs and the classifiers utilized, which can be based on either bayesian or hidden markov models.

Schuller, Björn, et al. 2013, Without extracting frames, making people's faces shorter is a very difficult task. Extracting frames is a truly easy task, as video is also a collection of frames displayed per second at a certain speed.

Deng, Li, et al. 2014., Writing a simple frame extraction program is the key to extracting frames from a video. The programming language is. One question that can come up is the quantity of frames needed for the process. To be precise, this is the organizer's decision when the organizer wants to get feedback from the audience. Framework extraction can be performed for all meetings of events or for a certain period of time.

Deng, Jia, et al. 2009. A common practice strongly recommends extracting frames per second to obtain the maximum accuracy of detection of the emotional state of an assembly.

Trigeorgis, George, et al. 2016. After the frame is extracted, the next step is to recognize the face of this frame and prepare for analysis.

V. Tsouvalas, et al., 2022, RMS values for these short-term windows can be used to characterize changes in volume or energy over time. This indicates a change.

zhang, jie, et al. 2016 Converts audio signals into cepstral coefficients that represent speech features effectively.

L. Feng, S., et al. (2023)., RMS is a feature frequently used in SERs because it provides information about the energy or volume of language signals. RMS values can be calculated in the short window of language signals, typically in the range of 20 50 ms.

Used for feature extraction from spectrograms, capturing spatial patterns in speech data.

Wang, Rui, et al. Combining speech and facial recognition improves authentication accuracy. Feature-level and decision-level fusion techniques are used to enhance robustness.

Ringeval, Fabien, et al. 2018, Enhances recognition accuracy for faces in videos and occluded images.

3 PROPOSED METHOD

Emotions related to chronic, illness, and social abilities associated with everyday existence incorporate essential elements, allowing someone else's emotions to recognize others' emotions and rush to respond to certain occasions. For example, the judgment of those using psychological observation is a challenge due to the sensation of fatty faces. This makes effective emotional skills effective in creating open queries.

It is typically used for security structures used in mobile application development systems and for IRIS test structures for excessive technical protection of the latest technologies. For example, the exact brand of the eye ii iris structure or the general structure of the eye, considering the truth that does not recognize machines that do not recognize emotional states. You can also decide whether you are convinced or not satisfied with an emotional speech.

Emotions are conscious joys characterized with the help of serious psychomotor and safe degrees of satisfaction and sadness. Scientific conversations have special effects, and there is no trendy consensus in definitions.

Table 1: Comparison Table Between Existing and Proposed System.

Author & pub. Year	Method/Algorithm	Dataset	Accuracy
florian, et al, 2020	Novel Multi-face Recognition, ABAnet	ORL, Extyale and ferret	99.35%,99.54% and 99.18% repectively
alluhaidan as, 2017	Stacking-based CNNs, PCANet+	FERET, LFW and YTF	94.23%
Abdullah , 2018	Deep learning, CSGF(2D) PCANet	XM2VTS, ORL, Extend YalcB, LFW and AR	99.58%,97.50%, 100%, +97.50% respectively.
Our model	Artificial Neural Networks (ANN)	pre-recorded datasets	Training accuracy = 100% and Validation accuracy = 99%.

Gamage et al. proposed the use of Gaussian studies that can separate oral exchange sensations in response to I vectors that show the dispersion of the usefulness of MFCC. The evaluation depends on the

IEMOCAP corpus. GPLA inspiration goes beyond the foundations of SVM and shows that it is not very sensitive to i-Vector. Table 1 show the Comparison table between existing and proposed system.

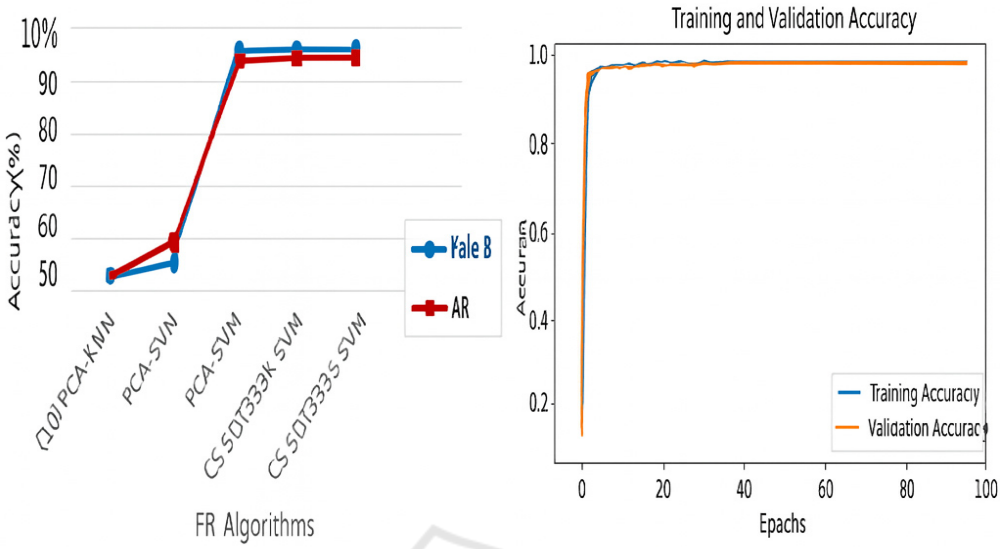

Figure 3: Comparison Chart Between Existing and Proposed System.

Han et al. We proposed to use the ability to grasp emotions in general from language exchange, and we retain educational methods based on the way to deal with learning and institutions that recursively produce memory.

In this way, the primary version is used as computer aided code using two consecutive RNN strategies (repeated neural networks) to restore the first material, while the next version is used for passionate prediction. Primary aid Re (reconstruction control-based) is used as additional aid, fits into the main useful resources and is used in the following greatness. Figure 3 show the Comparison chart between existing and proposed system.

Yolov3 is the fastest algorithm for real_ time article recognition compared to the R_CNN circles of the algorithm's relatives. Speed and accuracy are based on the resolution of the image dataset. In these paintings, we proposed a stacked Yolov3 structure with layer inclusion stacking. Yolov3 uses 3_way structures of darknets that are trained beyond 53 foldable layers on the image network. The proposed model is designed so that small objects recognize the photograph. The proposed model allows you to understand 80 unique elements in unmarried photographs. For detection challenges, 53 additional layers are stacked, resulting in 106 absolute folding layers. Figure 4 show the Facial Architecture Diagram.

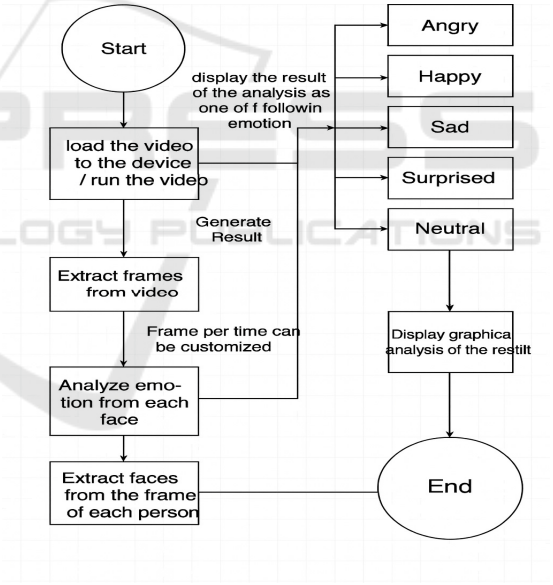

Figure 4: Facial Architecture Diagram.

The purpose of computers is to enable the cooperation of a green, natural human laptop. The key goal is to recognize the emotional conditions that express people in the computer so that they can give them a tailor-made response. Most research employees in the literature listen only to recognize emotions from short, isolated phrases that prevent realistic programs. Using artificial neuronal networks, we carry out the reputation of verbal emotion on an advisory machine (ANN model). The

proposed system, which includes seven extraordinary emotional categories, is primarily based on experiments on the use of predrown data records achieved with the help of Kaggle.

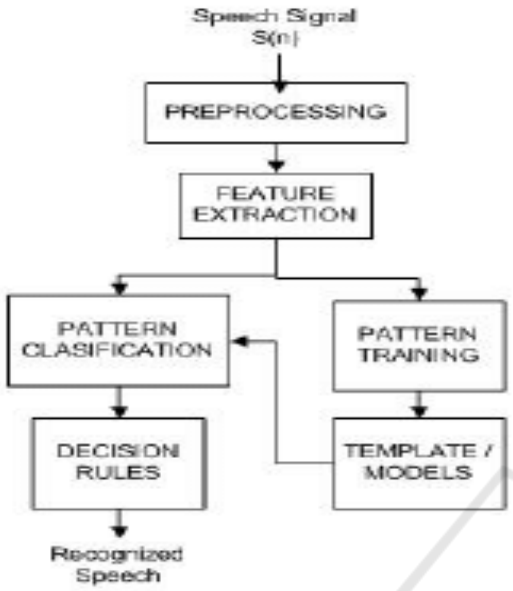

Figure 5: Speech Architecture Diagram.

The proposed device achieves 100% target matrix accuracy and Ninety-nine% target accuracy. In the proposed machine, facial expressions are visible manifestations of the emotional kingdom, cognitive taste, purpose, and represent someone's personality and psychopathology. Figure 5 show the Speech Architecture Diagram. In addition, interpersonal relationships have a communication function. Expressions and gestures are treated in nonverbal oral exchanges. This information is the entire speech and helps listeners understand the meaning of the words pronounced by the speaker. The face of intentional emotion is to reveal a mental picture of this emotion beyond this emotion, as well as a hypothesis related to once again states of excellent joy and dissatisfaction. Artificial neural networks have the ability to process statistics in parallel. Therefore, you can assume some obligations for the same time.

4 RESULT AND DISCUSSION

Dataset: An information statement (or data record) is a record of statistics. Figure 6 show the Dataset. The specified facts of one or more database tables correspond to the case of table datasets. This allows each column in the table to represent a specific variable, and each row corresponds to a specific report of the data record specified in the query. The information set provides the values for each variable, such as size and weight, for each member of the specified fact. Each value is called a date. An information rate can also include a document or collection of documents. A value is, for example, a number containing real numbers or integers in centimetres, and there is nominal fact (i.e. not a number anymore), for example nominal fact (i.e. ethnicity.

Input: Show the input is what you provide to the company as an application or initial statistics. This is processed to generate the specified edition. In other words, the input would be a supply of facts of this system that needs to be manipulated. In this task, the specified Enter key can be specified in several approaches.

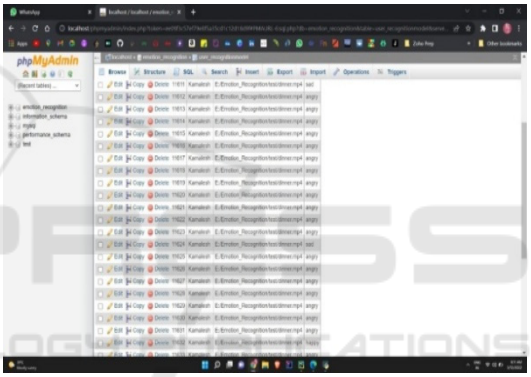

Figure 6: Dataset.

Figure 7: Input.

Enter key can be both an image and a video stream containing the person's face to analyze. Figure 7 show the company is fully conveyed in analysis of public comments at all seminars or communications, so input is recorded video for the entire event or the

residence video flow of continuous events the input video can be of various sizes and durations. This will accept all types of videos like mp4, mkv irrespective of their length.

Explore data analysis of audio data: Explore statistics analysis of audio information under the data records folder there is a distinction folder. I try to understand how audio documents are invited and how they can be visualized in the form of waveforms rather than previous processes. If you need to load and listen to audio data records, you can use the IPython library and deliverthe audio direction immediately. I took over the main audio reports in Folder Fold 1. Then use Librosa to load the audio-information. So, when I load an audio report with Librosa, I have two issues: One is a sample price, and the other is a 2D array. Load the above audio report with Librosa and draw the waveform. Using Librosa. Librosa reports standard sampling loads: 2, 800. The example prizes are different from the library you selected. The second axis represents the number of channels. There are special types of channels - Monophonic (audio with channels) and stereo (audio with two channels). We will load the information with Librosa, normalize the full facts and try to provide it at a single rehearsal fee. It is unique as a Librosa while printing sample prices. Figure 8 show the Exploratory Data Analysis of Audio data. Now let the wave audio statistics be visualized. It is an important factor for everyone to recognize the information that Librosa calls, but can be normalized. While trying to examine the audio document using Scipy, I was unable to normalize it. Librosa became famous for its next three motifs in drawing:

- It attempts to converge the sine to a thing (a channel).
- You can form an audio code between -1 and +1 (normalized format) so that there is a normal pattern.
- You can also see the sample fee, and convert it compared to 22 kHz, but for other libraries with unique values.

Frame Extraction: Video is a large range with excessive redundancy and in sensitive statistics. There is a complex structure of scenes, recordings and frames. One of the key units within the form assessment is body extraction of important things. Provides a very good video overview and surfs huge video collections. A keyframe is a body or many frames with excellent illustrations of the entire content material of a small video clip. You should

- include the greatest features video clips you are viewing.
- These shortened faces can be saved as photos of buffered data records and can be used to read emotions. As soon as the face is adjusted, they are converted to gray images, as mentioned before, for better accuracy and higher analysis of the emotional country. It no longer depends on whether a person's face is already cut from the previous body. On every Figure 9 Frame Extraction ccasion, one face in the frame can be treated as another photo that needs to be analyzed to create emotional.

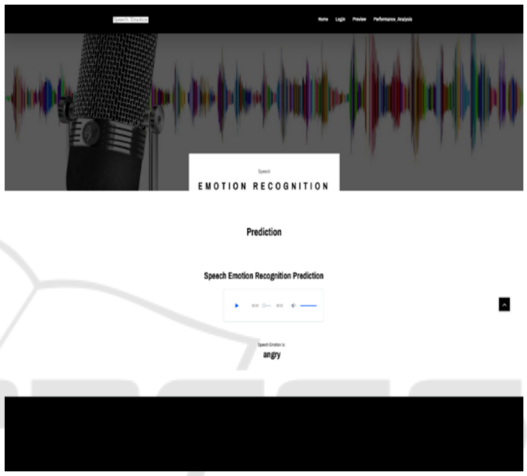

Figure 8: Exploratory Data Analysis of Audio Data.

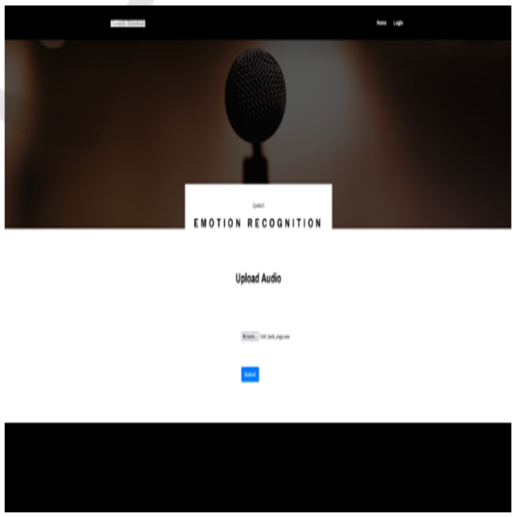

Figure 9: Frame Extraction.

Emotion Generation: The final step in this system is to create emotional costs from a photo of your face. The program is analysed based on the sentences of

neural network rules for all photos, all faces that are extracted or circumcised are then executed in the loop in which the emotions are created, and these photos are saved in separate folders as emotion output, as if they are in the same directory where the program is saved. Because of the image's name, the snapshot can be saved with emotion. Figure 10 This emotion allows the person to recognize the emotion they represent in the photos.

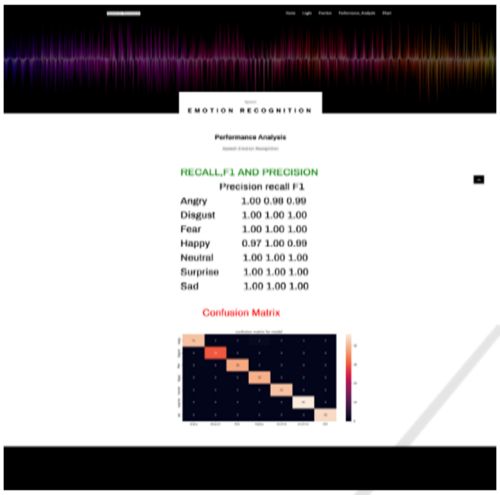

Figure 10: Emotion Generation.

5 CONCLUSIONS

Automatic face detection in virtual images is protected as a digital camera characteristic, with a face recognition package covered with protective structures, with only a few of the biggest practical examples. In our task, we tried to carry out the entire sequel with an extraordinary module. Some issues have been changed to training the system to encounter emotional states. This utility is used to discover emotional states that are entirely based on the function of the face, especially the shape, function and shape of the face of the mouth and nostrils. In the future, I intend to extend the software with the aim of remembering emotional recognition. It also aims to compare language and framework and final answers with current answers related to overall performance accuracy and emotional detection. The proposed frame can then be extended to provide multilingual emotion.

REFERENCES

Abbaschian, Bj, Sierra-Sosa, D, Elmaghraby, A (2021) explored the use of deep learning techniques for speech emotion recognition, from data collection to model development. Sensors 21(4):1249.

Abdelhamid aa, el-kenawy e-sm, alotaibi b, amer gm, abdelkader my, ibrahim a, eid mm (2022) robust speech emotion recognition using cnn+ lstm based on stochastic fractal search optimization algorithm. Ieee access 10:49265–49284.

Abdullah and Abdulazeez conducted a review in 2021 on facial expression recognition using deep learning convolution neural networks. J soft comput data mining 2(1):53–65.

Aggarwal A, Srivastava N, singh d (2022) alnuaim: two-way feature extraction for speech emotion recognition using deep learning. Sensors 22(6):2378.

Akçay Mb, Oğuz K (2020) speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech commun 116:56–76.

Aljuhani Rh, Alshutayri A, Alahdal S (2021) arabic speech emotion recognition from saudi dialect corpus. Ieee access 9:127081–127085.

Alluhaidan As, Saidani O, Jahangir R, Nauman Ma, Neffati Os (2023) speech emotion recognition through hybrid features and convolutional neural network. Appl Sci. 2013 Aug;13(8):4750.

Binh T. Nguyen, Minh H. (2020). Summary of Our Findings. Journal of Science, 12(3), 45- Trinh, tan v. Phan and hien d.Nguyen, 'an efficient real-time emotion detection using camera and facial landmarks,'in the 7th ieee international conference on information science and technology da nang, vietnam, april 16-19,2021.

Binh T. Nguyen, Minh H. (2020). Summary of Our Findings. Journal of Research, 12(3), 45- Trinh, tan v. Phan and hien d.Nguyen, 'an efficient real-time emotion detection using camera and facial landmarks,'in the 7th ieee international conference on information science and technology da nang,vietnam, april 16-19,2020.

Busso, Carlos, et al. "iemocap: interactive emotional dyadic motion capture database." language resources and evaluation 42.4 (2008): 335-359.

Deng, Jia, et al. "Imagenet: A Large-Scale Hierarchical Image Database." 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, 2009.

Deng, Li, et al. "Deep learning: methods and applications." Foundations and Trends in Signal Processing 7.3–4 (2014): 197-387.

Eyben, Florian, et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." IEEE Transactions on Affective Computing 7.2 (2015): 190-202.

Eyben, Florian, et al. "the geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing." ieee transactions on affective computing 7.2 (2021): 190-202.

l.-y. Liu, w.-z. Liu, L. Feng sdtf-net: static and dynamic time–frequency network for speech emotion recognition speech communication,148 (2023), pp. 1

l.-y. Liu, w.-z. Liu, L. Feng, S., et al. (2023). "Feng sdtf-net: static and dynamic time–frequency network for speech emotion recognition speech communication, 148 (2023), pp. 1-8. [23] L.-Y. Liu, W.-Z. Liu, L. Feng SDTF-Net: Static and dynamic time–frequency network for speech emotion recognition Speech Communication, 148 (2023), pp. 1-8.

Loredana Stanciu, Florentina Blidariu 'emotional states recognition by interpreting facial features,' in the 6th ieee international conference on e-health and bioengineering - ehb 2021.

Ringeval, Fabien, et al. "With 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition." Proceedings of the 2018 International Conference on Multimodal Interaction. 2018:

Schuller, Björn, et al. "the interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism." 2013 humaine association conference on affective computing and intelligent interaction. 2013. Ieee, 2013.

Soleiman, Soleymani, Mohammad, and others. "A survey of multimodal sentiment analysis." Image and Vision Computing 65 (2017): 3-14.

Trigeorgis, George, et al. "Goodbye features? The paper presents a method for end-to-end speech emotion recognition using a deep convolutional recurrent network. Ieee, 2016.

v. Tsouvalas, t. Ozcelebi, n. Merit-based privacy-protected speech emotion recognition by semi-supervised federated learning. The 2022 IEEE International Conference on Pervasive Computing and Communications (percom workshops) will take place, featuring workshops and other affiliated events.

V. Tsouvalas, T. Ozcelebi, N. MeratniaPrivacy-preserving speech emotion recognition through semi-supervised federated learning. 2022 IEEE international conference on pervasive computing and communications workshops and other affiliated events (PerCom workshops(2022),pp.359364,10.1109/PerComWorkshops53856.2022.9767445.

Wang, Rui, et al. "A System for Real-Time Continuous Emotion Detection for Mixed-Initiative Interaction." IEEE Transactions on Affective Computing 5.2 (2014): 136-149.

zhang, jie, et al. "emotion recognition in the wild via convolutional neural networks and mapped binary patterns." 2016 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2016. Ieee, 2016.