Prediction of Emotions Videos with Speech, Facial Expression Using MFCC-CNN

Sameena[1*], E. Srinivasulu[2]

Department of Electronics and Communication Engineering, Marri Laxman Reddy Institute of Technology and management, Dundigal, Hyderabad, India,

Email: sameena4128@gmail.com[1*], esasreenu@gmail.com

Abstract: Rapid progress in computer vision and machine learning has allowed for impressive achievements in fields such as object categorization, activity detection, and face recognition in recent years. Recent research and development in these areas have allowed for these victories. Despite this, identifying human emotions remains a formidable challenge. In recent days, a lot of work has gone into trying to figure out how to fix this issue. The advancement of artificial intelligence, natural language modelling systems, and related technologies has allowed for increased precision in this response to a variety of voice and speech-based methods. As a result of its generalizability, the study of emotions has the potential to contribute to many fields. One such area is working in tandem with human computers. Customers may get insight into their feelings, make more well-informed choices, and interact more naturally with robots with the help of computers. Predicting dynamic facial emotion expressions in cinema has received a lot of interest in recent years. About 10 years ago, this pattern first emerged. This study's authors propose a deep convolutional neural networks (CNNs) model for improving the accuracy and efficiency of emotion prediction using audio clips, still images of faces, and moving images. The mel-frequency Cepstrum coefficients (MFCC) are also recovered as a feature from the speech samples supplied by the speech CNN model. The given MFCC-CNN model outperformed baseline models in the end.

Keywords: Facial emotion Mel Frequency Cepstral Coefficient, convolutional neural network, speech emotion recognition

1.0 INTRODUCTION

In the context of social interactions, faces are one of the most essential visual stimuli because of the wealth of information they provide. This [1] includes one's identity, race, sex, physical appearance, and emotions. The human face is one of the most common visual cues. Early in development, humans have a high affinity for visually suggestive face traits and combinations. [2] According to studies, infants and even unborn children favor simple groupings that resemble faces. This predilection exists because it helps infants' bond with their caretakers and stimulates them to respond. Infants' responses in social interactions may be adaptable depending on the social partner's signals. Babies, for example, quickly learn to mimic their caregivers' facial expressions and vocalizations. Adults and school-aged children [3,] as well as older children, show a preference for certain facial visual traits. For instance, it may be easier to see a change in a person's facial features than in the features of an inanimate object, and it may take less time and effort to identify a person's face if it has been covered up than it would to do so with a mask.

It is possible to deduce the intentions and states of mind of others based on one's own thoughts and feelings via the use of two of the most important movable components of the face: gaze shifts

and facial expressions. We may concentrate on certain areas or things by shifting our gaze [4]. Some people's attention may be read by following the movement of their eyes. There are several considerations that go into figuring out the direction that someone else's eyes are gazing. While the iris-to-sclera ratio, head position, the existence of an item at the fixation location, and so on are all relevant, the emotional facial expression is the most important for the goals of this research.

You may learn a lot about a person's motivations and mental condition just by watching their faces. Furthermore, there is growing information that shows how the brain processes eye contact and how it impacts communication of emotion.

Multiple studies point to eye direction as a key factor in how long it takes to interpret a person's facial emotions. If someone is staring you down instead of avoiding their eyes, you have a better chance of figuring out whether they're angry or joyful. But as one looks away, grief and dread become more obvious. Adams and Kleck's signal-theoretic explanation for these findings provides a unified framework. According to this idea, pleasant emotions like pleasure and rage are 'approach-oriented,' whereas negative emotions like sorrow and fear are 'avoidance-oriented'.

However, it may be difficult to precisely judge the gaze of others around you if your own emotions are clouding your judgment. For instance, in research conducted in 2011, Lobmaier and Perrett had participants determine whether or not the depicted faces were staring squarely in their direction. The faces were shown in a variety of settings, each with its own lighting and backdrop. They came to the conclusion that smiles are seen to be directed towards the observer more often than other emotions such as fear, anger, or neutrality. The observations have been interpreted as opposing the shared signal theory considering the "self-referential positivity bias" hypothesis [5], which claims that individuals are more inclined to believe that they are the source of the enjoyment of others in order to increase their own feeling of self-esteem.

One of the most noticeable deviations in social behavior associated with autism spectrum disorder (ASD) is a diminished interest in human faces and failure in the face-related attentional systems [6]. People with autism spectrum disorder (ASD) have been demonstrated in a number of studies to pay disproportionately more attention to non-social than social stimuli. Individuals on the autism spectrum have been shown to have impaired face recognition memory and visual attention to facial signals. More evidence from the field of neuroscience suggests that those with ASD may struggle to read emotions conveyed via facial expressions. Reduced social bias and an aberrant orientation to faces, for example, have been uncovered in recent research that combine EEG and eye-tracking data. Studies were conducted on human participants diagnosed with ASD. The ordinary person pays more attention to their eyes than those with ASD, according to eye-tracking research. People with ASD, in contrast to TD, showed more activation of the social brain network when they diverted their gaze rather than when they gazed directly at another person. It seems to have more of an effect on how well gaze intentionality can be inferred than on how well glance direction or object identification can be understood.

Individuals with ASD have been found in several studies to have the same ability as TD individuals to correctly identify the direction in which someone is staring [7]. On the contrary, they seem to provide obstacles for integrating gaze orientation with communication and social settings. It was found that autistic children had difficulty making the connection between a person's facial expression and the direction in which they were gazing. This finding is relevant to the current investigation since it was found that children with autism had difficulty with it.

Multiple research' results have also led to the conclusion that ASD is the most severe form of a group of qualities related to social-emotional and communicative competence that are common in the population at large. Initial study findings on the association between the severity of autistic symptoms and performance on behavioral tasks diagnostic of ASD are presented by the Autism Spectrum Quotient (AQ). Examples of such activities include reading facial expressions to infer emotions, using eye contact to direct attention, and making assumptions about a person's mental state (Baron-Cohen et al., 2001a).

The primary purpose of this research was to determine whether those with high autistic features also consider facial expression while evaluating eye-gaze direction. The goal was to have a better understanding of autistic people's linguistic abilities. A person's gaze may tell you a lot about their motivations and attitudes, in addition to their manner. Studies have indicated that people on the autistic spectrum have difficulty encoding and integrating this information. It is debatable, however, whether this deficit is intrinsic to autistic features or rather the result of years of accumulated social experience shaping alternate patterns of communication. The most suitable participants for a study of the special contribution of autistic features would be typically functioning individuals who do not shun social connections. This would reduce the impact of demographic variables on academic performance, such as age and amount of social participation.

Caadas and Lupiáez (2012) [10] created a gaze discrimination task that we utilized to investigate the significance of eye-gaze direction in spatial interference paradigms. The brain processes information about a lateralized face more quickly and correctly when the face's glance is directed inward, toward the center of the scene. These authors backed up their assertion with supporting evidence. This effect was interpreted in terms of eye contact whereas classical results are typically observed with non-social stimuli like arrows. Additional studies support the "shared signal hypothesis" [11], which states that the inward effect is amplified when combined with approach-oriented emotions like happiness and anger and diminished when combined with avoidance-oriented emotions like fear. In-text citation: [space] Adding more "approach" emotions like happiness or rage heightened it.

The projections made by this research are as simple as they get. If the quantity of autistic features in the typical population is connected to these challenges, then only those with high levels should have difficulty integrating gaze direction with communicative and social situations. Low-autistic people shouldn't have these issues. People with high autistic qualities should be able to identify eye contact regardless of how they are feeling, whereas those with low autistic traits should be taught to control the inward impact of gaze direction. This was investigated in a study named "Autistic Traits and Gaze Direction."

Furthermore, there are two scenarios to consider when thinking about how facial emotions affect the detection of gaze control in people with minimal autistic symptoms. According to the shared signal theory, happy and angry expressions should have a larger impact on gaze direction identification than sad and scared ones (avoidance-oriented emotions). This would be the case if the emotions on people's faces genuinely affected where they looked. But if the "self-referential-positivity bias hypothesis" [12] is correct and emotional expression impacts the ID of gaze direction, then cheerful faces should have more internal effect than faces expressing other emotions or a neutral expression.

2.0 LITERATURE SURVEY

Zhang, Wei, et al. [13] created a unified modifier-based multimodal frame for identifying action units and recognizing facial expressions. To begin encoding the static vision feature, the active frame's image is utilized. Three distinct multimodal features are extracted from the video, audio, and textual content using a sliding window to clip adjacent frames. They then provide a fusion module based on a transformer to integrate the static visual qualities with the dynamic multimodal elements. Future detection jobs will be made easier with the help of a cross-attention element included into the fusion module, which directs the output participated features to zero in on the most important aspects of the fusion. They also use specific techniques for post-processing, data balance, and data augmentation to further enhance the model's performance. Their model dominated the EXPR and AU divisions in the official ABAW3 Competition. Ablation experiments and extensive quantitative assessments on the Aff-Wild2 dataset show that their method is successful.

Fard, Ali Pourramezan, et al. [14] proposed using an Adaptive Correlation (Ad-Corre) Loss to train the network in a manner that provides embedding feature vectors that are heavily connected with data from the same class but less so with data from other classes. Feature classifiers, mean classifiers, and embedding classifiers are used by Ad-Corre. The Feature Discriminator was created to aid the network in creating embedded feature vectors. This explains why embedded feature vectors of the same class a high degree of correlation with one another has, but feature vectors of different classes have a lower degree of correlation. The network's Mean Discriminator also produces an unnecessarily big disparity between the median embedded feature vectors of the various classes. The Xception network serves as the model's foundation, and they depart from earlier work by proposing an embedding feature space of k feature vectors. The Embedding Discriminator module will punish the network for not providing unique embedded feature vectors. They trained their model using the cross-entropy loss and the Ad-Corre Loss. Their recognition efforts on AffectNet, RAF-DB, and FER-2013 have shown excellent results. Extensive testing and ablation experiments demonstrate that their technique has the capacity to handle complex FER issues in the real world. Github hosts the primary folder for the code.

Bisogni, Carmen, et al. [15] studied the multiresolution face photographs' linguistic content. The last section uses transfer learning, progressive picture scaling, data augmentation, and fine-tuning parameters to extract more discriminating features and enhance the proposed system. The trials use the cohn-kanade dataset, the wild-face-expressions-static dataset, and the karolinska-directed-emotional-faces dataset, as well as a variety of existing methodologies for analyzing these datasets. When compared to other databases, the proposed solution clearly stands out as the winner.

Yu, Wenmeng, et al. [16] introduced a novel full-stack Co-attentive Multi-task Convolutional Neural Network (CMCNN). The Channel Co-Attention Module (CCAM) and the Spatial Co-Attention Module (SCAM) are the two nodes that make up this network. While performing FER and FLD tasks, the CCAM is responsible for determining channel co-attention scores by locating relationships between channels. SCAM spatial co-attention ratings employ a combination of max-pooling and average-pooling. Finally, they analyze the latest RAF, SFEW2, CK+, and Oulu-CASIA facial expression datasets. Extensive experimental data reveals that their approach outperforms single-task and multi-task baselines, proving multi-task learning's effectiveness and generalizability1.

Du Shichuan et al. [17] the findings of these studies are summarized in along with data suggesting that the same facial articulations used to make expressions in the lab are also used in real life. They also explore several unanswered research topics and comment on the implications of their findings for the study of mental disorders.

Shen, Junge, et al. [18] provide a new paradigm for evaluating student participation in class by having them utilize facial expression recognition to gauge their emotional states rapidly and precisely during the course. In addition, they offer a novel domain adaption-based approach to facial emotion detection that is especially useful in the context of MOOCs. The pilot findings validate their suggested framework as a reliable method for measuring students' engagement in their own learning. The results of comparisons with cutting-edge methods also reveal that their proposed technique is better for identifying facial expressions.

Xiao, Huafei, et al. [19] developed FERDERnet, an on-road driver emotion identification network based on facial expressions. This approach uses a deep convolutional neural network pre-trained on the FER and CK+ datasets and tuned for driver emotion recognition to break down the task of recognizing an on-road driver's emotions into three parts: face detection, augmentation-based resampling, and emotion recognition. Five backbone networks augment the ensemble technique. Additionally, to assess the suggested method, the researchers in this study created a dataset of drivers' facial expressions while they were really driving. This collection includes photographs of drivers in a range of situations, together with their facial expressions now. The dataset of drivers' facial expressions while driving was used in the experiments presented in this article. In a study that compared many state-of-the-art and baseline networks for their capacity to distinguish on-road drivers' facial expressions, the suggested FERDERnet with an Xception backbone achieved the highest results. The system's efficiency and accuracy were fundamental to the achievement of its goals.

Pabba, Chakradhar, et al. [20] made a real-time method for tracking student group involvement by studying facial expressions and identifying academic emotional states like "bored," "confused," "focused," "frustrated," "yawning," and "sleepy," which are important in the classroom. Face recognition and frame-by-frame calculations of group involvement are only two examples of the post- and pre-processing methods used into the approach. The face recognition algorithm is based on convolutional neural networks (CNNs). They used lecture recordings to construct a database of facial expressions to use in training the CNN model. They extended the model's predictions using illustrative data from the publicly accessible BAUM-1, DAiSEE, and YawDD datasets. Although the accuracy on the train was 78.70%, the accuracy on the exam was just 76.90%. There was a good correlation between the outcomes attained with the suggested strategy and students' ratings of their degree of involvement.

Andrey V. Savchenko et al. [21] of worried or cheerful expressions, and a move when neutral expressions were seen. Participants in the alternative version were shown the identical photographs but were instructed to respond differently depending on the gender of the actor shown rather than the emotions displayed on their faces. When the pleasant emotions served a goal, they discovered that they were more likely to diminish inhibitory control than negative emotions like dread. Based on their examination of these data, they conclude that a person's facial expressions do not always predict their conduct. Instead, this would occur only if necessary to accomplish the tasks at hand.

Umer, Saiyed, et al. [22] extracted face features using a single EfficientNet model pre-trained on the AffectNet database, they developed a unique frame-level technique for emotion identification. That's why they think their approach might be applied to other fields, such mobile video analytics. They show that their basic model significantly outperforms the VggFace baseline using experimental findings from the Aff-Wild2 database used in the third Affective Behavior Analysis in the Wild (ABAW) Competition. On the Expression Classification validation sets, and Expression Classification, Valence-Arousal Estimation, their method outperforms the state-of-the-art by a factor

of 0.15 to 0.20. Since their method can be applied to all four sub-challenges with little effort, it may soon replace the current gold standard.

Saurav, Sumeet, et al. [23] four-part technique is explained below. The first step of the system involves using the input picture to carry out a process known as face detection. To perform feature learning tasks for classification, a convolutional neural network architecture trained using deep learning has been built in the second stage. The goal was to improve the chances of effectively extracting relevant and distinguishing features. Finally, the face image has undergone several novel data augmentation techniques to boost the learning parameters of the proposed CNN model. This was done to make the developing system more effective generally. The fourth section is a method for refining the trained CNN model, which requires choosing between amassing more data and using deep learning features. GENKI-4k (2 expression classes), KDEF (7 expression classes), and CK+ (7 expression classes) are three reference databases used to demonstrate the breadth of the experimental outcomes. They have shown and extensively examined the suggested system's performance about each database, and they have compared these outcomes to those of current methodologies that are presently considered state-of-the-art. The suggested system is shown to be superior to current approaches.

Fang, Zheng, et al. [24] developed a robust DICNN model for in-the-wild facial expression identification in real time. This model was built with the goal of being used in an embedded environment. The created DICNN model achieves maximum performance with just 1.08 million parameters and 5.40 MB of RAM. To do this, they must strike a balance between the accuracy of the recognition and the speed of the computation. They evaluated the DICNN model against state-of-the-art approaches on four different FER benchmark datasets (FERPlus, RAF-DB, CKPlus, and FER2013,) utilizing a wide range of presentation evaluation metrics. Accuracy, precision, recall, and F1-score were some of the quality indicators used for recognition. The improved DICNN model was then put on an Nvidia Xavier embedded platform, making it highly transportable and inferrable at high speeds. To do this, the TensorRT Software Development Kit was used. The proposed FER system outperformed previous systems widely considered to be state-of-the-art in terms of execution speed improvement and accuracy, respectively.

Wang, Kai, et al. [25] developed a simple but effective Self-Cure Network (SCN) to reduce the number of open questions. For clarity, SCN utilizes a self-attention mechanism on the FER dataset to weight each training sample with a ranking regularization and a painstaking relabeling process to alter the labels of the lowest-ranked group. In the FER data collection process, both approaches are employed. Their approach has been validated by experiments on both simulated FER datasets and the real-world WebEmotion dataset. Their SCN outperforms the state-of-the-art by a wide margin, as seen by its performance on public benchmarks (88.14 % on RAF-DB, 60.3 % on AffectNet, and 89.5 % on FERPlus).

3.0 PROPOSED METHOD

We present a full-stack deep learning system built on an attentional convolutional network for analyzing face images to deduce the underlying emotion. This framework allows for the possible recognition of facial emotions like smiling and frowning. Standard methods to improve the performance of a deep neural network include adding additional layers and neurons, increasing the network's regularizations (by, for example, implementing spectral normalization), etc. This is particularly the case for projects that need many categories for categorization. Despite this, we show that, because to the limited number of classes, a convolutional network with less than 10 layers and

attention (which is trained from scratch) may outperform state-of-the-art models for facial emotion recognition. This is shown by demonstrating that the network is taught from the top down.

Oftentimes, we just need to concentrate on select sections of an image of a face to get a sense of the underlying mood represented by the snapshot, even if we can't see all of the person's facial features. In response to this discovery, we included a spatial transformer network into our framework to serve as an attention mechanism for focusing on key aspects of the face.

In Figure 1 we see the architecture of the suggested paradigm. In this architecture, a max-pooling layer separates each group of convolutional layers, and a rectified linear unit (ReLU) is used as the activation function. A dropout layer happened after two complete ones. Two convolution layers, max-pooling, ReLU, and two fully linked layers made up the localization network (or spatial transformer).. Initially, the input was transformed into the sample grid T (θ), and then the transformation parameters were regressed to produce the distorted data. To focus on the parts of the picture that are most important for the task at hand, the spatial transformer module estimates a sample over the area of interest. In this case, we employed an affine transformation, which is one of several possible transformations that may be used to introduce distortion into the result. new information about the spatial transformator apparatus.

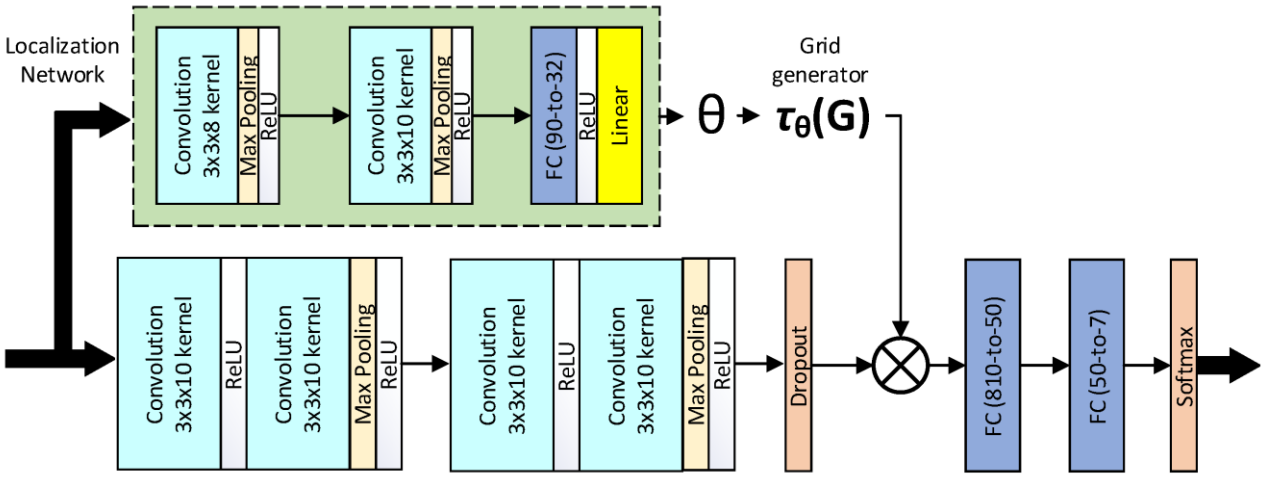

Figure 1. The proposed model architecture.

After that, we used stochastic gradient descent (in particular, the Adam optimizer) to fine-tune a loss function during model training. The classification loss (also known as cross-entropy) is added to the l2 norm of the weights in the last two fully connected layers to get the loss function.

$$L_{overall} = L_{classifier} + \lambda \left\| w_{(fc)} \right\|_2^2 \qquad (1)$$

Based on the model's confirmation set performance, the regularization weight (λ) is tweaked to find the ideal value for optimal performance. We train our models from abandon on tiny datasets like JAFFE and CK+ with the aid of dropout and l2 regularization. This is made possible by combining these two regularizations. It is important to note that we trained a fresh model for each dataset used in this study. Using a network with a roughly comparable architecture but with more than fifty layers did not greatly increase accuracy, so we abandoned that approach. We found that the network with fewer layers was better in inference speed and was a good fit for real-time applications.

A. DATASET DESCRIPTION

The development of a BORIS project file simplified the research process. This document was an ethogram for the labels that would be applied to the primary video clips during annotation. Each annotator was given a total of 218 files, including the movie you're now watching. An individual BORIS project file was created for each video's annotations. When we were through annotating, we centralized all of the data.

Annotations made by different researchers may be compared using specialized software. This application combines and merges several output files that include information about parts with the same labels. For an annotation to be considered credible, it must be agreed upon by two of the three researchers participating in the study.

To finish off the dataset, we extracted emotional tagged segments from the input videos and saved them as separate output files. For this objective, a software based on the ffmpeg library was created. For the sake of this investigation, the following presumptions were made:

- Pieces of coverage noted by two or three observers are merged into a single, triangulated fragment marked by all three.
- Each fragment was given a time buffer of 1.5 seconds when counting from the beginning of the recording, and 0.5 seconds when counting from the conclusion of the recording.
- We fixed the length of the remaining unlabeled items to be equal to the sum of all labels for the topic, and we categorized them as "neutral" based on the total number of labels for the subject.
- Those pieces that were labeled "no face" have been thrown out.

When the script was run, video files saved in Matroska Media Container (MKV) format with either the h264/opus or vp8/opus source encoding were created. Clips vary in duration from 6.3 seconds to about 23 seconds, on average. Information about the dataset itself may be found in a file named fragments.csv, which is included in the dataset's download. This file describes the dataset and contains details like the dataset's ID, job name, label, duration, filename, note observers, video codec, and all observers. The whole of the package was compacted, and the resulting zip archive is now available as a separate file.

B. PREPROCESSING

Preprocessing face image data may help remove the effects of factors like head position, lighting, and occlusion (such as glasses, facial hair, or self-occlusion) on facial emotions. As a result, there is substantial variation in the results obtained by neural networks when analyzing various contextual manifestations of facial emotion. Then, thorough preprocessing may enhance the accuracy of face emotion recognition.

Using the MTCNN technique, we identified the five anatomical features of the face (the eyes, the nose, and the two sides of the mouth) across all pictures in the facial expression dataset [49]. Photos of facial expressions were aligned after certain adjustments were made based on their similarities. The facial expression photos were then normalized by having their dimensions standardized to 224 by 224 pixels. We conducted a five-fold cross-validation test on the CK+ and Oulu-CASIA datasets. Our network was trained using a validation set created from a subset of the original dataset that was split into five equal sections. Accuracy levels were obtained by taking an average of five different tests in each category. We give both the accuracy and the average accuracy of the dataset due to the asymmetry between the categories in the RAF-DB dataset (the average

accuracy is the average of the sum of all category accuracies). The reliability of the data set is improved by include a training set in addition to a test set.

C. MFCC

The first step in creating a speech recognition system is extracting features from the audio signal to determine which parts of the signal are best at recognizing the linguistic content and which parts are worst at recognizing other types of information conveyed by the signal (such as background noise, emotion, and so on). The system now separates out the signal components most indicative of language use. One of the most common techniques for extracting features is shown in Figure 2; this is the Mel-frequency cepstral coefficients technique. The audio route may be completely contained inside the current power spectrum. Occasionally, MFCCs are used to stand in for this sphere. In the 1980s, Davis and Mermelstein promoted the use of MFCCs for use in automated speech and speaker identification.

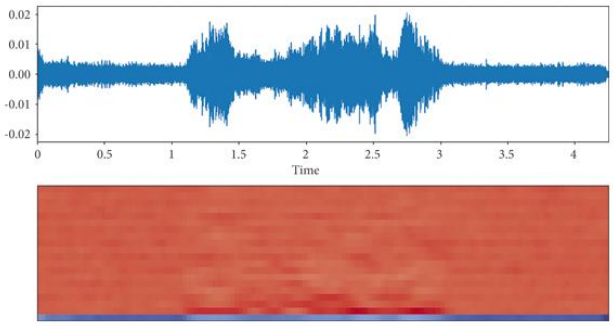

Figure .2. MFCC extraction for happy sound data in RAVDESS dataset.

The Mel-frequency cepstrum illustrates the short-term power spectrum of a sound by using the linear cosine transform of a log power spectrum presented on a nonlinear Mel-frequency scale. The cepstral representation of the audio sample is given by the Mel-frequency cepstral coefficients. When compared to a typical cepstrum, a Mel-frequency cepstrum has frequency bands spaced in a regular fashion that corresponds to the Mel scale. As a result, the ear's reaction becomes more consistent with that of the discrete frequency bands seen in a typical spectrum.

D. CNN MODELS

Google has trained Inception-v3 on the large ImageNet picture collection, making it ready to be used immediately for image classification applications. More than 5 billion instructions multiplying and adding together the model's approximately 25 million components would be required to categorize a single image. The Inception-v3 model can quickly establish an image's categorization even on high-end desktop PCs without a GPU. The 15 million images in ImageNet's archive have been split up into 22,000 categories. One million photographs from 1,000 different categories make up the LSVRC, the most prestigious picture classification competition taking place right now, and their subset is identical to that. It is not practical to train the deep model on a standard desktop PC, since doing so may take several weeks. Instead, a regular computer can only be used to train a shallow model. In this study, the pretrained Inception-v3 model is used to classify facial expressions. You may get this pre-trained model from the internet and use it to label images of face emotions. The logic flow for the CNN model is shown in Figure 3 for your perusal.

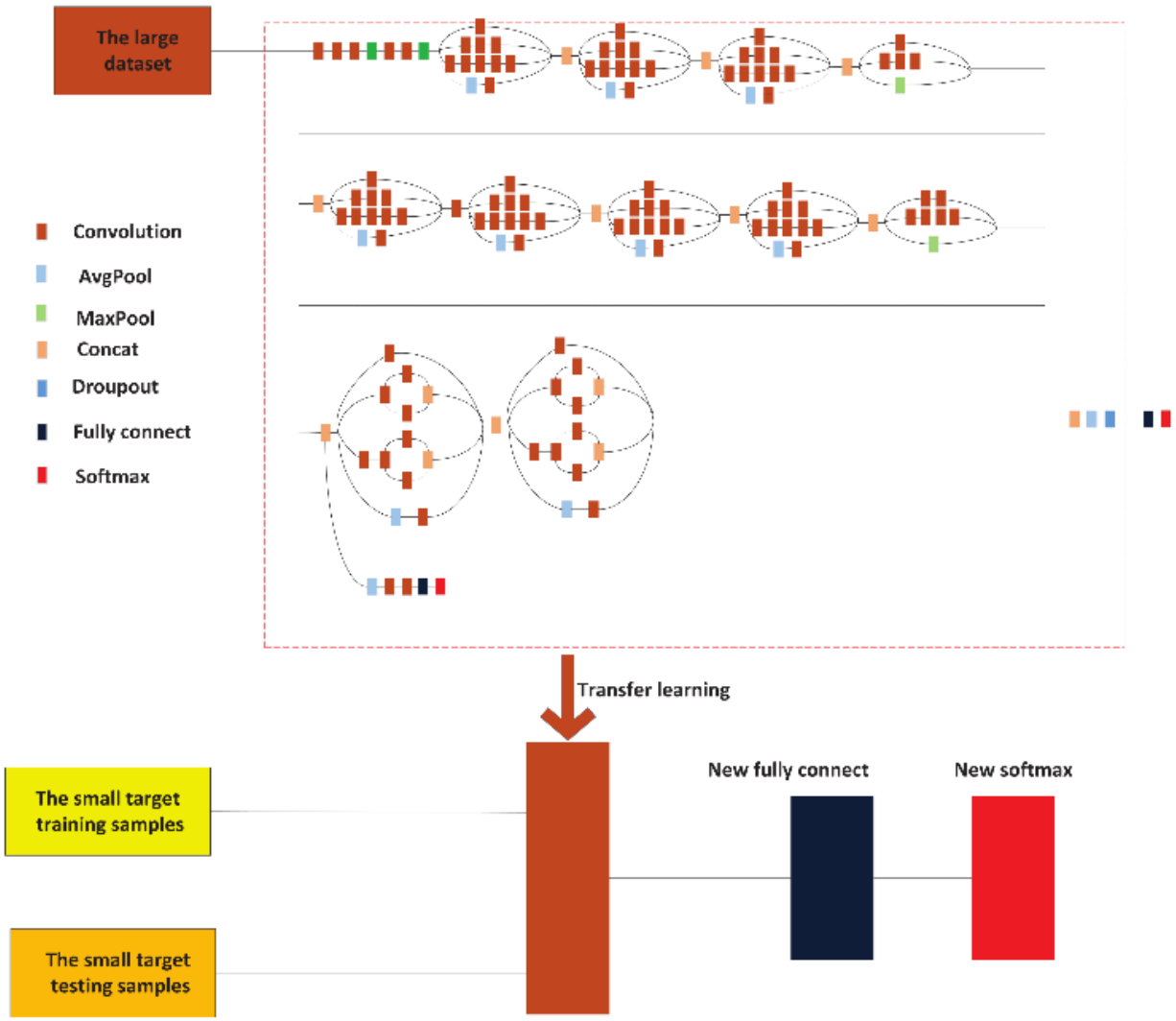

Figure 3. A flowchart illustrating the transfer learning process. The new entirely connect layer and the new softmax layer are retrained using the new target training data, and these parameters may be quickly transferred to the small new training samples for fine-tuning. The red dotted line represents the parameters that were trained in the big sample.

4.0 RESULTS AND DISCUSSIONS

The full simulation study findings are shown here. Test pictures for facial expression-based emotion prediction are shown in Figure 4. The included emotions are shock, disgust, awe, rage, and fear. The loss performance and acquired prediction accuracy of the recommended deep CNN for face expression, audio, and video are shown in Figures 5 and 6, respectively. The suggested deep CNN clearly excelled both facial expression and voice inputs when it comes to predicting emotions from videos, as shown by the two graphs. The relative merits of the many suggested CNN models and the already available methodologies are summarized in Table 1. When compared to state-of-the-art ANN, RNN, and LSTM models, the deep CNN suggested here significantly outperformed the competition.

Fig. 4: Sample test images of emotion prediction.

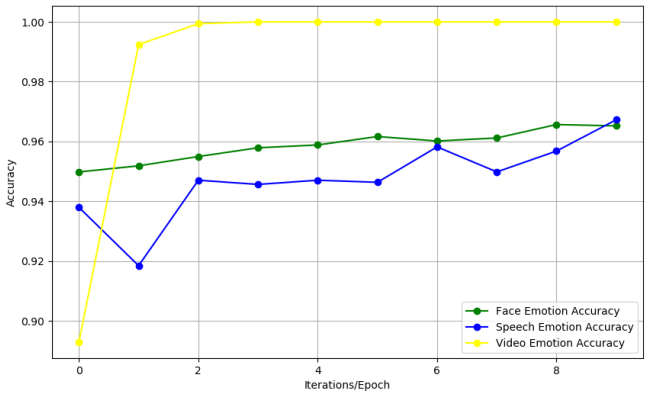

Fig. 5: Performance comparison of prediction accuracy using proposed deep CNN with speech, facial expression, and videos.

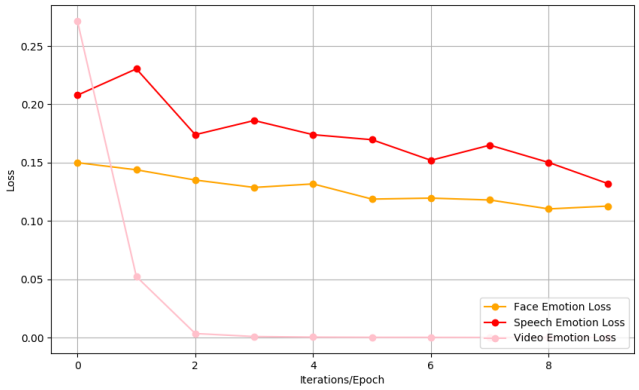

Fig. 6: Performance comparison of prediction loss using proposed deep CNN with speech, facial expression and videos.

Table 1. Accuracy performance comparison

Dataset	**ANN [13]**	**RNN [15]**	**LSTM [17]**	**Proposed Deep CNN**
Facial expression	76.26%	81.23%	90.345%	96%
Speech expression	87.34%	89.34%	92.345%	97%
Video expression	91.26%	93.45%	95.78%	100%

5.0 CONCLUSION

Sentiment analysis has grown in prominence as a field of study in recent years due to the range and depth of information it may give for various applications. No matter how much we try to hide it, our true selves come out in the words we pick and the expressions we create. A wide variety of information, such as but not limited to spoken, written, and visual data, may be used in emotional interpretation. As a result, the authors of this research developed a deep convolutional neural network (CNN) model with improved prediction accuracy and reduced loss for identifying emotions in spoken language, facial expressions, and video. The MFCC was also used to extract features from the many voice samples used to train the speech CNN model. It's possible that expanding this training to incorporate more facial expressions may be required to obtain even higher levels of performance.

References

1. Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." *Sensors* 21.4 (2021): 1249.

2. Kwon, Soonil. "Optimal feature selection based speech emotion recognition using two- stream deep convolutional neural network." *International Journal of Intelligent Systems* 36, no. 9 (2021): 5116-5135.

3. Avots, E., Sapiński, T., Bachmann, M., & Kamińska, D. (2019). Audiovisual emotion recognition in wild. Machine Vision and Applications, 30(5), 975-985.

4. Wang, Xusheng, Xing Chen, and Congjun Cao. "Human emotion recognition by optimally fusing facial expression and speech feature." *Signal Processing: Image Communication* 84 (2020): 115831.

5. Kerkeni, Leila, et al. "Automatic speech emotion recognition using machine learning." *Social media and machine learning*. IntechOpen, 2019.

6. Pandey, Sandeep Kumar, Hanumant Singh Shekhawat, and SR Mahadeva Prasanna. "Deep learning techniques for speech emotion recognition: A review." *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2019.

7. Özseven, Turgut. "Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition." *Applied Acoustics* 142 (2018): 70-77.

8. Tarunika, K., R. B. Pradeeba, and P. Aruna. "Applying machine learning techniques for speech emotion recognition." *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2018.

9. Schoneveld, Liam, Alice Othmani, and Hazem Abdelkawy. "Leveraging recent advances in deep learning for audio-visual emotion recognition." *Pattern Recognition Letters* 146 (2021): 1-7.

10. Vryzas, Nikolaos, et al. "Continuous speech emotion recognition with convolutional neural networks." *Journal of the Audio Engineering Society* 68.1/2 (2020): 14-24.

11. Neumann, Michael. "Cross-lingual and multilingual speech emotion recognition on english and french." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

12. Hossain, M. Shamim, and Ghulam Muhammad. "Emotion recognition using deep learning approach from audio–visual emotional big data." *Information Fusion* 49 (2019): 69-78.

13. Pan, Zexu, et al. "Multi-modal attention for speech emotion Zhang, Wei, et al. "Transformer-based multimodal information fusion for facial expression analysis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

14. Fard, Ali Pourramezan, and Mohammad H. Mahoor. "Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild." IEEE Access 10 (2022): 26756-26768.

15. Bisogni, Carmen, et al. "Impact of deep learning approaches on facial expression recognition in healthcare industries." IEEE Transactions on Industrial Informatics 18.8 (2022): 5619-5627.

16. Yu, Wenmeng, and Hua Xu. "Co-attentive multi-task convolutional neural network for facial expression recognition." Pattern Recognition 123 (2022): 108401.

17. Du, Shichuan, and Aleix M. Martinez. "Compound facial expressions of emotion: from basic research to clinical applications." Dialogues in clinical neuroscience (2022).

18. Shen, Junge, et al. "Assessing learning engagement based on facial expression recognition in MOOC's scenario." Multimedia Systems (2022): 1-10.

19. Xiao, Huafei, et al. "On-road driver emotion recognition using facial expression." Applied Sciences 12.2 (2022): 807.

20. Pabba, Chakradhar, and Praveen Kumar. "An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition." Expert Systems 39.1 (2022): e12839.

21. Savchenko, Andrey V. "Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices." arXiv preprint arXiv:2203.13436 (2022).

22. Umer, Saiyed, et al. "Facial expression recognition with trade-offs between data augmentation and deep learning features." Journal of Ambient Intelligence and Humanized Computing (2022): 1-15.

23. Saurav, Sumeet, et al. "Dual integrated convolutional neural network for real-time facial expression recognition in the wild." The Visual Computer (2022): 1-14.

24. Fang, Zheng, et al. "Facial expression GAN for voice-driven face generation." The Visual Computer (2022): 1-14.

25. Wang, Kai, et al. "Suppressing uncertainties for large-scale facial expression recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.