Predicting Evoked Expression from Videos Using Convolutional LSTM

MSc Research Project
Data Analytics

Kishore Kumar Anandhapadmanaban
Student ID: X21130850

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: .....Kishore Kumar Anandhapadmanaban.....

Student ID: .....X21130850.....

Programme: .....Data Analytics..... **Year:** 2022-2023

Module: ..... MSc Research Project .....

Supervisor: ..... Teerath Kumar Meghwar.....

Submission Due Date: .....18/09/2023.....

Project Title: Predicting Evoked Expression from Video using Convolutional LSTM

Word Count: .....8588..... **Page Count**.....24.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....Kishore Kumar Anandhapadmanaban.....

Date: .....18/09/2023.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□✓

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Evoked Expression from Videos using Convolutional LSTM

Kishore Kumar Anandhapadmanban
X21130850
National College of Ireland

Abstract

Evoked expressions in videos, representing spontaneous emotional reactions, are fundamental to understanding human emotions and intentions. Predicting these expressions poses significant challenges due to the intricate interplay of spatial, temporal, and auditory aspects in videos. Traditional methods often rely on unimodal strategies, utilizing either visual or auditory cues, and might employ conventional model architectures Convolutional Neural Network (CNN) or Visual Geometry Group (VGG-16), which often fail to capture the multifaceted nature of expressions. In this research the data is from the Evoked Expressions from Videos (EEV) dataset, the study involves several diligent stages. The collection of pictures from videos is followed by the production of spectrograms from the matching audio, resulting in an extensive visualization of data. This research introduces a novel multimodal approach, amalgamating audio, and visual cues, and implements a Convolutional Long-Short Term Memory (Conv-LSTM) model to capitalize on both spatial and temporal dimensions. Our approach overcomes existing limitations by integrating the spatial feature extraction capabilities of CNNs with the sequential modelling strengths of Long Short-Term Memory (LSTM). Other than that, our research compares the performance of our proposed model with Custom CNN model and VGG-16 algorithm. Rigorous evaluations demonstrate that our model outperforms existing methods in terms of validation loss, Mean Square Error, and MeanAbsolute Error, offering a robust and effective solution for evoked expression prediction from videos. The Mean absolute error of the Conv-LSTM is 0.197 and Mean Square error is 0.064 and evaluation on validation loss is 0.64. The Conv-LSTM overperforms the other two models.

Keywords: Evoked Expression, EVV dataset, CNN, Conv-LSTM, VGG-16, Mean Square Error, Mean Absolute Error and Validation Error.

1. Introduction

Emotions can be described as the inherent characteristics reflected by humans and deliberately appearing as voluntary or involuntary aspects depending on facial expressions. These expressions are mainly recognized when individuals come in contact with a flow of communication, visual

effect and audio. Considering typical non-verbal communication, facial expressions are indicative of a suitable phase of analyzing human emotions. In fact, research studies have vividly drawn attention to the aspects in various domains to understand both macro and micro-expressions perceived through human emotions. According to Ben *et al.*, (2022), a broad classification of facial expressions is mainly classified under macro and micro-expressions wherein the main difference is reflected in two significant classes implicating durations and intensities. When focusing on macro-expressions, the demonstration depicts a voluntary expression which covers a major portion of the facial area. Correspondingly, micro-expression served knowledge on the involuntary movement of the facial area - a rapid as well as local expression. It is because of certain inherent properties; micro-expressions are often difficult to identify through the naked eye except for experts trained in reading little to non-existent facial expressions. Apparently, human-based analysis of different micro-expressions is time-consuming, error-prone and expensive. Therefore, ideal methods are necessary to adopt, which are automatic and developing a computer-aided vision as well as identifying patterns of these facial expressions.

While understanding the significance and approach to understanding facial expressions, the source of this expression is equally essential to explore. Videos are often specified on the basis of semantics as well as effective content. Herein, the semantic content values “What exactly is present in the video?” and the effective content represents “What feelings does the video happen to make out of people”. Research has significantly imposed an understanding to consider and explore evoked expressions based on videos and audio effects from those videos. The explicit patterns are often studied by researchers to annotate audience expressions typically evoked because of the content. Usually, datasets such as “Evoked Expressions from Videos” (EEV) are used to purposively study the impact of these evoked expressions. From a large-scale implication and understanding of the experience evaluated for viewers, it has been observed that visual as well as auditory information of viewers can necessarily evoke vivid responses. In that case, the automatic prediction of these evoked expressions based on videos promotes an enhancement and engagement between recommended systems and different social machines for better interaction with users. Typically, a renowned large-scale dataset named “Evoked Expression in Videos” or EEV dataset is used to study viewers' responses which are usually derived from the visual and auditory effects.

In the prediction of these evoked emotions, researchers necessarily use distinguished features and algorithms that can depict the intended and experienced emotions. Wang *et al.,* (2021) in their study has precisely described the approach and focused on predicting audiences’ emotions based on a purposive exploration of the “valence-arouse scale” over a particular time being. Oftentimes, in this prediction, it has been observed that some methods must gather image features and optically enhanced information from different videos and movies which, therefore, imposes extensive computational complexities. Studies in recent times have provided promising results in the context of describing videos based on semantic content and parallelly worked on understanding the evoking enhanced by these videos on viewers. While using semantic content, datasets that are extremely responsive to provide promising results include Sorts1M, Kinetics and YouTube8M and hence enabled a prominent understanding of semantic videos. Apparently, the current approach to studying datasets for effective responses to video content is extremely scarce. Therefore, this challenge is another contributing factor which limited the collection of framed-

level effective videos. Moreover, Sun *et al.,* (2020) implored that it is considerably difficult to identify reliable post-evoke stimuli as well as effective labels in a subjective manner considering the background and context of viewers.

In speculation to construct a large-scale dataset to understand affect-based content, different approaches are enhanced by researchers in their studies wherein various algorithms as part of scalable methods are significantly used in annotating facial expressions. However, several challenges in analyzing the critical expressions from video content have limited the abilities of conventional models. Over the years, the prediction of intended or complex experienced emotions is explored on a valence-arousal scale. However, the drawback presented with the prediction process by the conventional models seems to observe the unaccounted information on temporal structures wherein expression changes evoked through audio-visual information are completely ignored. Thus, it increased computational challenges regarding feature extraction, and therefore, proceeded to explore new methods which can mitigate these challenges.

1.1 Aim of Research

The aim of this study is to advance the prediction of evoked expressions from videos by integrating audio and visual cues. Through the implementation of a Conv-LSTM model, we endeavor to seamlessly harness both spatial and temporal dimensions, striving to outperform existing methodologies in both accuracy and robustness.

1.2 Research Objectives

- Objective 1: To investigate the effectiveness of combining audio and visual cues from videos in amultimodal approach, with the goal of enhancing the accuracy and robustness in predicting evokedexpressions.

- Objective 2: To design, implement, and compare various model architectures, including custom CNNs, VGG-16, and a Conv-LSTM model, focusing on their ability to handle the spatial and temporal dynamics inherent in videos for expression recognition.

- Objective 3: To conduct a rigorous quantitative assessment of the proposed models using metricslike Mean Square Error, Mean Absolute Error, and validation loss, aiming to identify the optimalmodel and reveal insights into its strengths and areas for potential improvement.

1.3 Research Question

To navigate this complex landscape and uncover the underlying mechanics of expression recognition, the following research questions have been formulated.

- In predicting evoked expressions from videos, what are the relative performances of different model architectures, such as custom CNNs, VGG-16, and Conv-LSTM? How do these models manage the spatial and temporal complexities inherent in video data and which model generates the most accurate results?

2. Literature Review

2.1 Chapter overview

The chapter provides information on face recognition and the classification process wherein emphasis on different models highlights the significance of the models in predicting the evoked expression of humans. The specific knowledge acquired from this chapter typically represents a range of indications in the automatic prediction of evoked expressions based on the visual and audio effect of videos. The implication of findings from the literature is further evaluated by analyzing the application of datasets preferably EEV, which is extensively used by researchers and experts to study viewer responses toward videos. In this observation, the implication based on facial expressions is annotated through machines and interpretations are further validated by testing both human and machine-based annotations. In this chapter, all this information has been discussed by amplifying insightful knowledge on different models - their accuracy and limitations in the prediction process.

2.2 Evoked Expressions from Videos (EEV)

Videos have been observed to evoke different responses from viewers. While acknowledging this fact, it is imperative that the ability to identify the evoked expression of the audience prior to their watching can necessarily help in the development of the content as well as video recommendation. In this regard, studies introduced a large-scale dataset, named Evoked Expression from Videos or EEV to study viewers' responses (GitHub, 2023). Coming to an understanding of the composition of this dataset, EEV typically consists of 3 different files - "train.csv", "val.csv" and "test.csv": thus representing the training, validation & test files splits accordingly. Each of these CSV files indicates different yet expected facial expressions implicated by individuals that have reacted to certain videos, specifically to the visual and audio effects. Among different lines, the initial line within CSV indicates the header with labels representing each column.

Research presented in previous studies implicated that due to the huge size of the dataset, EEV consists of approximately 4.8 million face expressions annotations of viewers who reacted to nearly 18,500 videos (Sun *et al.,* 2020). As previously discussed, to consider and specify a comparable result, experts specifically verify the split files wherein the training file is totally machine-annotated however, the other two (validation and test files) have human & machine annotations. According to the performance obtained based on machine annotations, shows an average rate of precision of 73.3% respectively. Apart from this, various other approaches to acquiring improved results with effective content have been established over the years with existing models. Thus, this study has developed a comparable note from this discussion to identify which model provides an improved response with the EEV dataset.

2.3 Machine Learning-based Face Recognition

Facial expression-based emotion recognition (FEER) demonstrates intuitive reflection indicating an individual's mental state is enriched with emotional information. It is in fact, has been considered informative and important intrapersonal communication. Song, (2021) in the study provides a demonstrative view and knowledge of the complexity as well as variability in the facial expression of humans which oftentimes posed challenges in the prediction process. The author explained that traditional methods of evaluating these expressions have shown insufficiency due to a lack of prominent features and susceptible conditions because of the external influence of the environment. Over the years, different supervised machine learning approaches have been introduced in the prediction process.

Song, (2021) in their study introduced a "fusion dual-channel expression recognition" algorithm that was established through machine learning-based theory as well as philosophy. The features extracted and used are based on Gabor taken from the ROI area and typically used the convolutional neural network (CNN) without recognizing subtle changes that happened to be indicated in expressions. To be precise, the proposed machine learning algorithm has followed two different paths wherein the first path considered Gabor features to actively recognize the emotion area locally. Consequently, the second path introduces an efficient channel-based network, which consists of a depth-separated convolution for improving the straight bottleneck structure. Based on this approach, the complexity previously observed in human features is reduced and restricts overfitting through designing efficient modules of attention combining depth features based on spatial information. As per the identification of the performance, it has been recognized that the extracted Gabor features from the ROI area have provided improved recognition accuracyusing the CNN model and outperformed the competitive dataset FER2013.

The detection of human emotion under different approaches seeks additional security and information of the person. Based on different studies, it has been identified that human emotion recognition is a complex process which requires overwhelming effort in facial expression-based emotion recognition. With the advent of understanding maximum to minimum expressions can

differ depending on the subtle changes that happen to be observed in the expression. Raut, (2019) in their study explained the fact that contortion in facial muscles at a minimum or insignificant level can be challenging to predict. Moreover, these expressions might vary among individuals since emotions are highly context dependent. Thus, focusing on this information, the author explained that extracting features and categorizing them primarily necessitates the approach to recognizing the specific model to be used for the purpose. In this regard, machine learning models such as Support Vector Machines (SVM) are used, and features are extracted from the human emotion's dataset. As per the understanding of the information on the dataset, the author implied that the dataset could enable a robust approach and depict the classification nature of the proposed algorithm (Raut, 2019). According to performance identification for the SVM algorithm, the model can provide an accurate identification process to recognize emotion although further research has been implied with more improved models to prevent challenges most certain in image processing methods of emotions.

With the advent of understanding the implication behind different facial expressions, research on "Facial Emotion Recognition" (FEM) is an extensive approach which is elucidated by experts in various domains. An effective analysis of facial emotion is often used for different purposes which, therefore, need proper specification and comparability with benchmarked results. In the study conducted by Khan, (2022), the author introduced different algorithms particularly machine learning and deep learning models that are used to predict human emotions from FER datasets. Evaluation metrics are used to compare the results and also explore gaps in these models which would be a criterion for future research implications. Considering the outcome obtained from the study, it has been observed that both conventional machine learning and deep learning models have provided improved emotion classification based on extracted features, however, the models are time-consuming both in the training phase and testing phase when demonstrating micro-expressions. Thus, it is increasingly difficult to achieve better performance with features of minimal facial expressions, especially due to various external environmental impacts. Hence, it is suggestive to explore improved models which can provide distinguished results for both macro and micro facial expressions of viewers.

2.4 Deep Learning-based Faced Recognition

The proper understandings of evoked emotions which are induced from certain visual and audio effects have been documented for years to study those expressions with utter interest. However, expressions which are timid and almost non-existent are difficult to decipher by machines. Ellis *et al.,* (2019) explained that both industry experts and academicians have greatly advanced to conduct sentiment & emotion analysis although most approaches are mainly resonated in the content domain. The author explained the intended work contribution by studying evoked emotions of viewers after viewing movie trailers. These emotional responses from movies and videos are considered to be well-suited to predict emotions from different expressions. However, it is equally addressed in some studies that apart from suitable recognition of emotions there is difficulty in

bridging the gap in “affective labels” thus, inflicting challenges in modeling the high-level human emotion based on low-level visual and audio features (Ellis *et al.,* 2019). Therefore, suitable models are used by experts and addressed in studies to undermine the advances in the prediction process.

Emotion recognition based on conversation is another important approach which widely gains attention in research. The widespread application based on emotion recognition is acknowledged in different sectors such as healthcare, education, and human resources and implicates the psychological domain. To establish a suitable and different approach to the recognition process, Ghosal *et al.,* (2019) introduced a graphical neural network model, named “Dialogue-Graph Convolutional Neural Networks” (DG-CNN or Dialogue-GCNN). A real specification of the approach has leveraged self as well as inter-speaker dependency indicating interlocutors to model-based conversational contexts to predict the emotion. As per the observation of the performance, it is identified that the model properly addressed issues related to context propagation and certainly outperformed previous models with comparable benchmarked results based on classification datasets. This approach although specified conversational-based emotion recognition rather than video-based evoked expression; the understanding provides an implication that deep neural networks performed better than conventional machine learning models. Another study conducted by Mellouk & Handouzi, (2020) has explained a perspective view of the automatic facial expression approach gaining interest in recent times. Pertaining to the information reviewed from different studies, it is vividly considered that emotion prediction is a rather contemporary research attention which is implicit to gain information on a particular person as needed. Therefore, feature extraction and accurate prediction of the emotion are enhanced by better modeling architectures such as deep neural networks. Mellouk & Handouzi, (2020) in the study elicit the use of deep learning architectures to study features extracted from FER datasets, whilst the result obtained is compared with other benchmarked results. From the overall evaluation, it is identifiable that the model is progressive to provide improved recognition accuracy; However, the paper explicitly guides researchers to future research implications to make further improvements.

Prediction of emotions that have been evoked from viewers while watching movies or short videos has become a consideration in research to analyze affective video content for different purposes. Usually, emotions evoked by audiences are combined with visual-audio effects which are further testified by extracting features from scribed datasets significantly used for prediction purposes. A research study conducted by Wang *et al.,* (2021) has used distinct audio and visual features that can help in the production of experienced emotions although combined semantic information that can refine these features to enhance the prediction result. As a suitable model can improve prediction accuracy, the long short-term memory (LSTM) model has been used by incorporating “temporal attention mechanisms” to predict experienced emotions. Another study conducted by Thi *et al.,* (2023) demonstrated the approach to develop and predict emotions through multi-modal models of viewers from video-based face expressions. A hybrid model of prediction has been

introduced depending on clips of videos and audio. Distinctly, two approaches are specified one of which classified movie clips through deep connected layers without any memory while another approach is based on the LSTM network. Based on the result obtained, it can be stated that the performance accuracy for emotions predicted independently is slightly better than with the LSTM model. Huynh *et al.,* (2021) demonstrated a similar approach to predict emotions. However, used EEV datasets to extract features and take benefits from pre-trained models. The model used is a "temporal convolutional neural network" which explored temporal relationships and gained advantages based on memory consumption as well as parallelism. As per the achieved result based on the model and EEV datasets, the coefficient obtained is 0.05477, showing a first-ranked performance of EEV in 2021.

2.5 Classification and Predictability through Different Approaches

Emotion influences an individual's well-being and most certainly depicts the interaction pattern, actions, and judgments. To understand human emotions, researchers have taken an interest in predicting sentiment based on recognizing human expression upon watching videos, through conversation and sometimes studying the music effect. The affective response of video or movie viewers as explained by Phuong Thao *et al.,* (2021) is helpful in the prediction of intended or experienced emotions. To suitably establish an accurate prediction result, different models are used by researchers, especially conventional machine learning models and some effective deep neural architecture. However, previously observed, even though those models have provided better results, these models are time-consuming and deemed unfit in the prediction of micro expressions from EEV datasets. Nguyen *et al.,* (2023) explained an approach to transformer architecture, which gained attention in the recent decade. It has become a dominating paradigm to study many applications including the affecting computing process. Therefore, to study the emotion of viewers based on the EEV dataset, transformer-based architectures can be used as an alternative method.

3. Methodology

In this study, the aim is to anticipate the feelings that a video clip will arouse at particular moments. Choosing the type of content that will be interesting to a certain client requires consideration of desired sentiments. This is particularly true when determining which video a spectator would like the best. In such scenarios, it is essential to extract the various expressions that a viewer expresses while streaming, which can be accessed by deploying the deep learning model which can extract the semantics from the images using highly complex structures therefore in this work deep learning algorithms based on the different domains are incorporated. Since forecasting using deep learning is not a straightforward process but consists of various intensive processes like data collection, data preprocessing, data analysis andvisualizations, feature engineering, initialization and training of algorithms followed by the

evaluation of trained algorithms utilizing different metrics to get a model which can be deployable in real-world applications. Since all these steps are performed in this study. Each process undertaken is explained in detail in further subsections and overall flow is described using methodology diagram as shown in Figure 1.

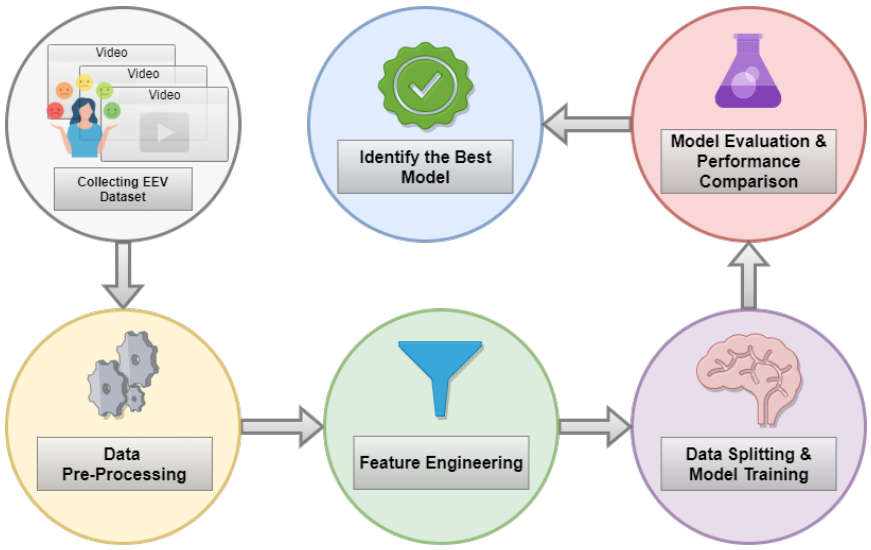

Figure 1 Methodology Flow Diagram for Predicting Evoked Expression from Videos

3.1 Dataset Description

To deploy deep learning techniques to identify and accurately classify facial utterances in a streaming video the most important step is to gather a rich and labelled dataset. A key tool for studying how viewers respond to videos is the Invoked Experiences through Videos (EEV) dataset, which makes it possible to examine how viewers' faces change as they react. The collection is extensive, containing 470 hours of video footage and nearly eight million observations from 5,154 clips. A wide range of sensations, including enjoyment, frustration, amazement, focus, misunderstandings, disapproval, satisfaction, dissatisfaction contradiction, joyfulness, fascination, discomfort, sorrow, shock, and glory, are captured in every clip at an oscillation rate of 6 Hz using 15 constantly evoked phrase labels. The degree of assurance that each expression of the face has been captured in that frame is indicated by the quantification of these expression labels, which range from 0 to 1. It is impossible to emphasize the importance of an enormous data set in the field of deep learning like EEV. Massive data sets have many benefits, but they are especially useful for building reliable and precise models. Potential worries about confidentiality of information, infringement of intellectual property, and data validity are substantially minimized by receiving the dataset through AIcrowd, a trusted platform. The integrity and validity of the study findings are also enhanced by the transparency and reliability of a reliable source, thus enhancing the task's worth.

3.2 Data Preprocessing

Every research work that utilizes data effort, particularly deep learning operations, must start with data preparation. Preparing raw data for analysis and modeling entails cleaning, converting, and organizing it. To build a condensed and significant depiction in this particular instance, it was crucial to extract pertinent frames and related sensations from the video footage. Additionally, by recording both aural and visual signals during the process of creating pictures with associated audio spectrograms, the dataset was enhanced by adding noise, providing a more thorough insight into audience responses as shown in Figure 2. The obtained Evoked Expressions extracted from the Videos (EEV) data undergo a number of crucial procedures throughout the data preparation phase to get it ready for additional analysis. First, several folders were created to organize the data, including "Train_Images" for the storage of video clip frames that were extracted. This data on provoked gestures was then read from a CSV file that included labels for 15 different facial expressions and their related confidence levels. Five major facial expressions were chosen in order to condense the dataset and concentrate on the most important information: concentration, bewilderment, contentment, curiosity, and melancholy. The 0.9 quantile of the collected data was determined. This decision highlighted the primary feelings of interest in the research and allowed for more manageable data processing. The next step of the data preparation procedure was to gather fragments from every clip. The images were recovered depending on the timestamps that corresponded to viewer facial expressions for each movie in the dataset's first 500 items. Only the frames that matched the chosen emotions were kept because they were linked to certain emotional expressions in the frames. To ensure the inclusion of highly expressive examples, frames with confidence scores in excess of or comparable to the highest ten percentile for each chosen emotion were also taken into account. A comparable picture, made up of the frame in question and a spectrogram taken from the audio of the video, was created for each eligible frame. The data was reorganized into a new CSV file for further applications.

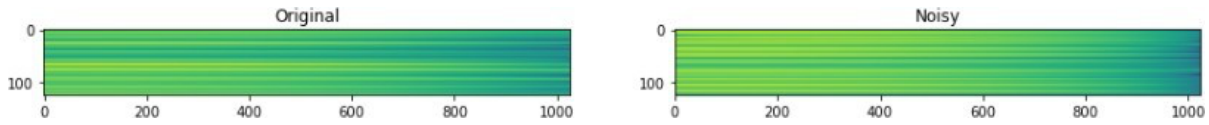

Figure 2 Data After adding Noise

3.3 Data Visualization

Exploratory Data Analysis focuses more on the features and structure of the image collection and also aids in the selection of data preprocessing and model architecture by looking at the dimensions, image quality, color channels, and spatial distribution of the images. EDA assists in choosing the best data augmentation approaches to expand the dataset's size and variety, which is necessary for efficiently developing algorithms using deep learning. In this work, some analysis and visualization of data is performed. First, the text data provided in the CSV format is analyzed by looking first and last five rows followed by the analysis of the expression’s data distribution as

shown in Figure 3. Further, the counts of each category in emotion columns are analyzed with the help of a bar plot as shown in Figure 4.

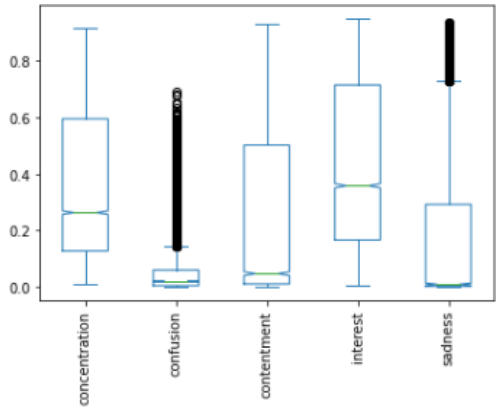

Figure 3 Data Distribution of Expressions

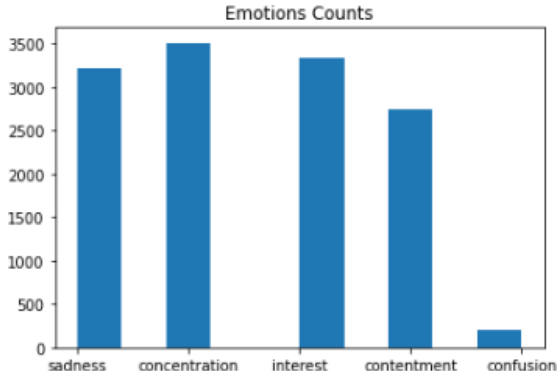

Figure 4 Counts of Expression in Data

In the next analysis, the sample of image data is visualized by plotting the scatter-polar plot of the associated label to the image as shown in Figures 5, 6, and 7. From Figure 5 it can be noted that the image contains some sort of concentration and more inclination toward contentment and interest while the image in Figure 6 inclines highly to sadness and some to interest and concentration. From Figure 5 it is clear that the image shows high contentment and low parts on interest.

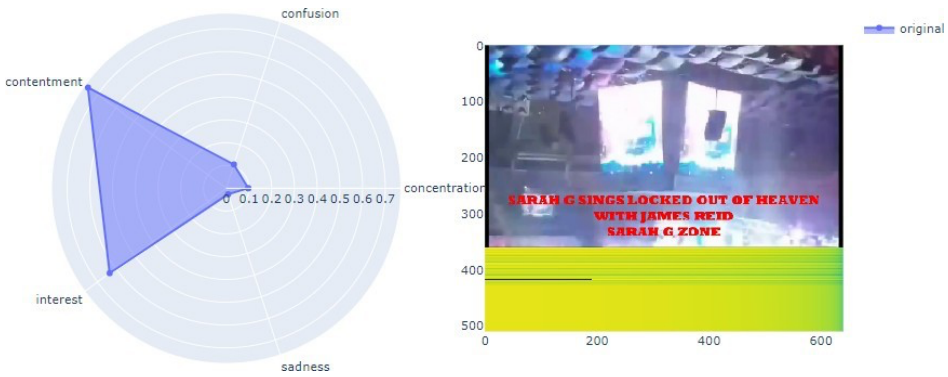

Figure 5 Visualization of Expression in Image and Audio

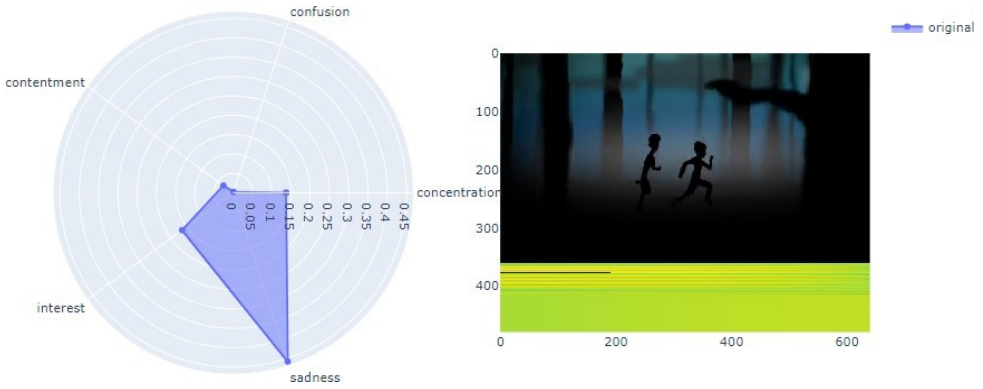

Figure 6 Visualization of Expression in Image and Audio

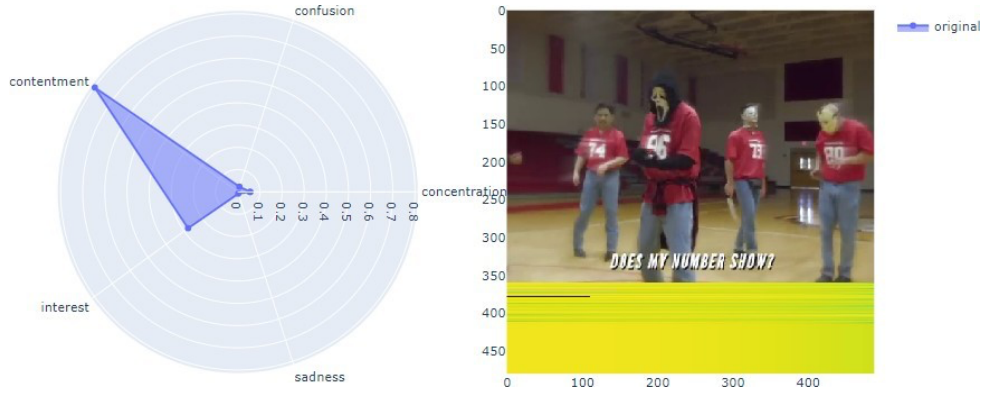

Figure 7 Visualization of Expression in Image and Audio

3.4 Feature Engineering

In this work, videos are utilized as a prime source of data while the extracted images and labels associated with each image are present in the CSV file, which arises the doubt of irregular frames and other uncertainties therefore, Prior to putting the raw data into the algorithm, image preprocessing aims to improve both the accuracy and the significance of the dataset. To maintain consistency across all images, the method retrieves the image first prior to resizing it to a given width and height of 640,480 with a channel size of 3. This scaling step is crucial because it uniformizes the picture proportions and eliminates any potential inconsistencies that might result from different initial sizes. The images are then normalized by scaling all pixel values between 0 and 1, enabling more rapid convergence during training, by reducing all component values by 255. Through the use of a generator function to process data in batches, the preprocessing procedure efficiently handles large datasets and enables iterative model training, enhancing the overall training speed and performance. Preparation of images makes ensuring that the model receives images that are uniform in size and format, preventing any possible problems brought on by

variable image dimensions. The effectiveness of the deep learning model in identifying and forecasting evoked emotions from films is substantially influenced by image preprocessing.

3.5 Model Training

Three deep learning models—a Customised CNN approach, a VGG-16 model (using transfer learning strategies), and a Conv-LSTM model—were used in this research work for the modeling phase to identify elicited emotions from videos. Despite changing the sequence, the data were divided into test and training sets in a ratio of 80 to 20. The VGG-16 algorithm model used pre-trained weights gained from an enormous image dataset to improve its capacity to generalize, while the customized CNN model was created to capture distinctive properties from the images recovered during preprocessing. The Conv-LSTM model, which combines Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs), was used to make use of the spatial and temporal information included in the video frames. A single batch of 4 was employed for the purpose of training data, whereas 1 was used for the validation data. The study explored several methods for emotion identification from video data using a combination of basic CNN, transfer learning using VGG-16, and the sophisticated Conv-LSTM model.

3.6 Model Evaluation

The model assessment gives us crucial information regarding the advantages and disadvantages of the various deep learning models used in this study, allowing us to assess their effectiveness and select the best model for expression identification from videos. Additionally, testing the models on a different dataset before judging them assures that the evaluation is fair and accurate of the models' capacity for handling unknown data. The customized CNN approach, the VGG-16 model, and the Conv-LSTM model were among the developed deep learning models whose efficacy was evaluated in the present study using the relevant measures. Mean Absolute Error (MAE), and Mean Squared Error (MSE), coupled with the loss function employed during model training, was selected as the main assessment measure due to the ongoing nature of the expression's values. This makes it possible to quantify how effectively the algorithms anticipate evoked responses from frames of video.

4. Design Specification

In this study, to carry out the application of deep learning algorithms a combination of different strategies offered by deep learning are implemented. Three distinctive deep learning algorithms are employed which are a Custom CNN Model, a Conv-LSTM model, and a Vgg-16 model which is a pretrained model and invokes the concept of transfer learning. The architecture of each model is described in the following subsections.

4.1 Convolutional Neural Network (Custom CNN) Model

The custom CNN architecture is designed to analyze and classify expressions within frames of a video. It begins with a Convolutional layer employing 32 filters of size 3x3, each using the ReLU activation function. Subsequently, a MaxPooling layer with a 2x2 pool size is employed to downsample the spatial dimensions and reduce computation. The architecture further deepens with another Convolutional layer, now utilizing 64 filters of the same 3x3 dimensions and ReLU activation. Following this, another MaxPooling layer is employed. The subsequent Flattening layer transforms the 2D feature maps into a 1D vector, allowing compatibility with traditional fully connected layers. A Dense layer comprising 128 neurons and employing the ReLU activation function follows, capturing high-level features in the data. For the final classification, the model concludes with an output Dense layer containing 5 neurons, corresponding to the 5 different emotions the model aims to classify. Notably, this layer utilizes a linear activation function, as the network is treated as a regression problem rather than a traditional classification task. The model is trained using the Adam optimizer with the mean squared error (MSE) loss function and Mean Absolute Error (MAE) as an evaluation metric. Architecture of Convolution neural network is shown in Figure 8.

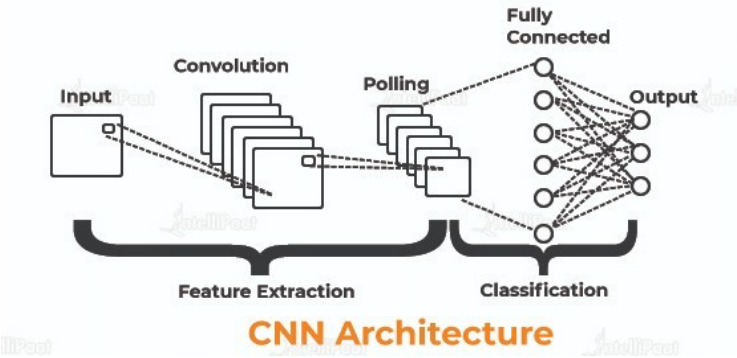

Figure 8 Convolutional Neural Network Model Phung. and Rhee. (2019)

4.2 Convolutional Long-Short-Term (Conv-LSTM) Model

The Conv-LSTM model is designed to process sequential image data while capturing spatial and temporal features effectively. It begins with a stack of convolutional layers to extract visual patterns. Convolutional filters of size 3x3 are applied, followed by a rectified linear unit (ReLU) activation to introduce non-linearity. Subsequent max-pooling layers with 2x2 filters reduce spatial dimensions while retaining essential information. This aids in learning hierarchical features. Another convolutional layer set with 64 filters is employed, enhancing feature abstraction. The architecture transitions from convolutional layers to a flattened operation, converting the 3D feature maps into a 1D vector. The LSTM layer is introduced to capture temporal dependencies, essential for understanding sequential patterns. With 128 units, the LSTM processes the sequence and outputs its hidden state. Following the LSTM, fully connected layers contribute further abstraction. A dense layer with ReLU activation aids in learning complex relationships. The final dense layer with five neurons and linear activation yield emotion predictions. The model is trained

using the Adam optimizer and optimized for mean squared error loss while monitoring the mean absolute error metric. Architecture of Conv-LSTM model is illustrated in Figure 9.

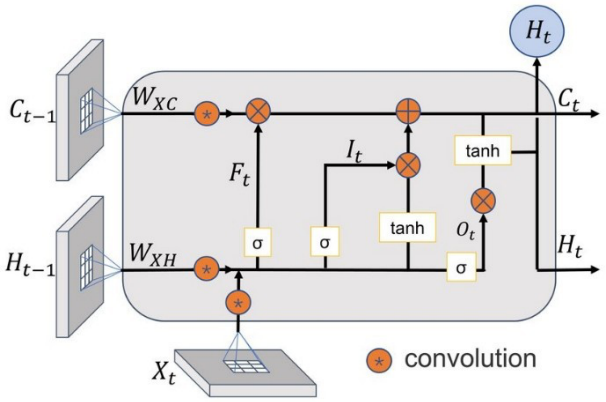

Figure 9 Convolutional Long-Short-Term Model Shi, C. et al. (2022)

4.3 Vgg-16 Model

The VGG16 model is a deep convolutional neural network architecture that gained prominence for its simplicity and effectiveness in image classification tasks. its deep structure is composed of 16 weight layers, including 13 convolutional layers and 3 fully connected layers. The architecture's key feature is the consistent use of 3x3 convolutional filters throughout the network, which contributes to its depth while maintaining a relatively small receptive field. VGG16's convolutional layers are organized in sequential blocks, with each block containing multiple convolutional layers followed by max-pooling layers for spatial downsampling. In transfer learning, VGG16's pretrained version on the ImageNet dataset is often utilized as a feature extractor for various tasks. By freezing the weights of the pretrained layers and adding custom layers on top, the model can be fine-tuned for specific tasks. The top layers are modified to adapt the network's output to the desired number of classes or regression targets. This approach leverages the learned features from ImageNet and allows the model to generalize effectively to new tasks while requiring fewer training samples. Architecture of VGG-16 model is illustrated in Figure 10.

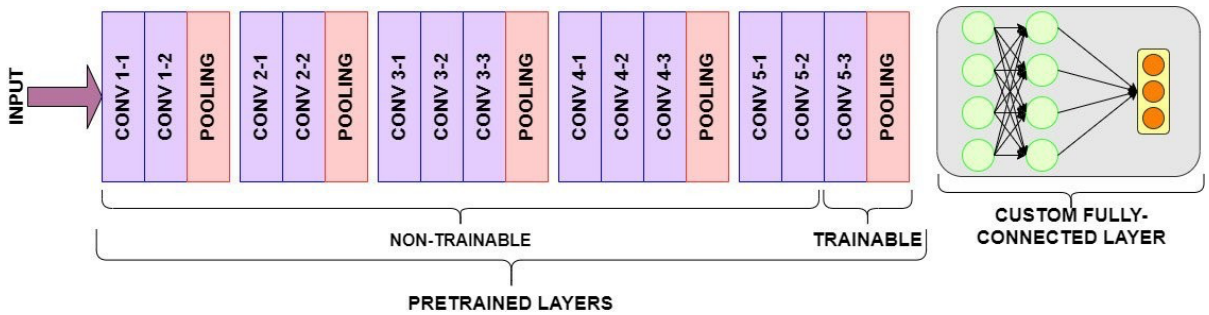

Figure 10 Vgg-16 Model James McDermott , (2021)

5. Implementation

This work's execution involves a number of crucial phases, all of which were carried out in the language used for programming, Python. Initially, the Evoked Actions via Clips (EEV) dataset, which consists of 8.1 million observations of audience expressive reactions to 5,154 images, was gathered from a reliable source, AIcrowd. Only five noteworthy expressions were chosen since the dataset only has 15 consecutive evoked expression labels. Pandas and Numpy were two Python modules used for data processing and modification. The preprocessing of the data was done after data collection in order to build a spectrum from the audio and identify frames from the movies. Images and videos were read and edited using the OpenCV library. For effective model training, the preprocessed pictures were subsequently scaled and normalized. For the purpose of optimizing pixel values and guaranteeing uniformity in picture dimensions, image preprocessing was crucial. The data generator function made it easier to input data during training in batches, which improved memory consumption effectiveness. The implementation used TensorFlow as the backend and the Keras library for deep learning models. Next, the dataset was split into training and test sets in an 80:20 ratio without shuffling to preserve temporal order. Three deep learning models were implemented for evoked expression recognition: a Custom CNN model, a VGG-16 model (leveraging transfer learning), and a Conv-LSTM model (integrating spatial and temporal information). Model training involved iterative epochs with forward and backward passes, and model parameters were updated using an optimizer. The performance of the models was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and loss metrics. The Scikit-learn library facilitated MAE and MSE calculation, and Keras' loss functions were utilized during training. The algorithms' prediction abilities and applicability for emotion identification from video frames had to be evaluated, which was crucial. A quantitative evaluation of the model's performance in foretelling elicited expressions was made possible by looking at MAE and loss measures. The assessment procedure made sure that the models would perform well on new data and be generalizable. The dataset's properties and the performance of the models were better understood through illuminating plots and charts created with Python's Matplotlib and Seaborn tools. The implementation in Python included libraries like Numpy, Pandas, Seaborn, Matplotlib, OpenCV, Librosa, VideoFileClip, Sklearn, Keras, and Tensorflow.

6. Result and Evaluation

After the training of each deployed algorithm on the training set of data, it is crucial to analyze and assess the obtained results on the test data because test data is analogous to the real-world unseen data of the algorithm. The algorithm which performs best on the test data is considered the superlative algorithm despite the fact that any other algorithm might be performed best on the training data. Therefore, in this work, three evaluation metrics which are MAE, MSE, and validation loss are equipped to evaluate the models on the testing data. The obtained values on these metrics for each algorithm are discussed in further sub-sections.

6.1 Evaluation based on Mean Absolute Error

Effective results were obtained from the evaluation of the deep learning algorithms' efficiency using the Mean Absolute Error (MAE) measure. Lower values of MAE indicate greater predictive performance, which is an indicator of the model's reliability in predicting elicited emotions from video frames. The Custom CNN model showed encouraging results when each model was trained for 25 epochs, yielding a validation MAE of 0.219. The Conv-LSTM model showed a validation MAE of 0.197. These findings demonstrate that the algorithm produced estimates that were highly precise and had a relatively low percentage of errors on the validation sets. The VGG-16 model produced a validation MAE of 0.198 by utilizing transfer learning. Which is very similar to the Validation MAE of Conv-LSTM model. The comparison of Algorithms based on validation MAE is shown in Figure 11. Formula of MAE can be derived as follow.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_1 - \hat{y}_i\right|^2$$

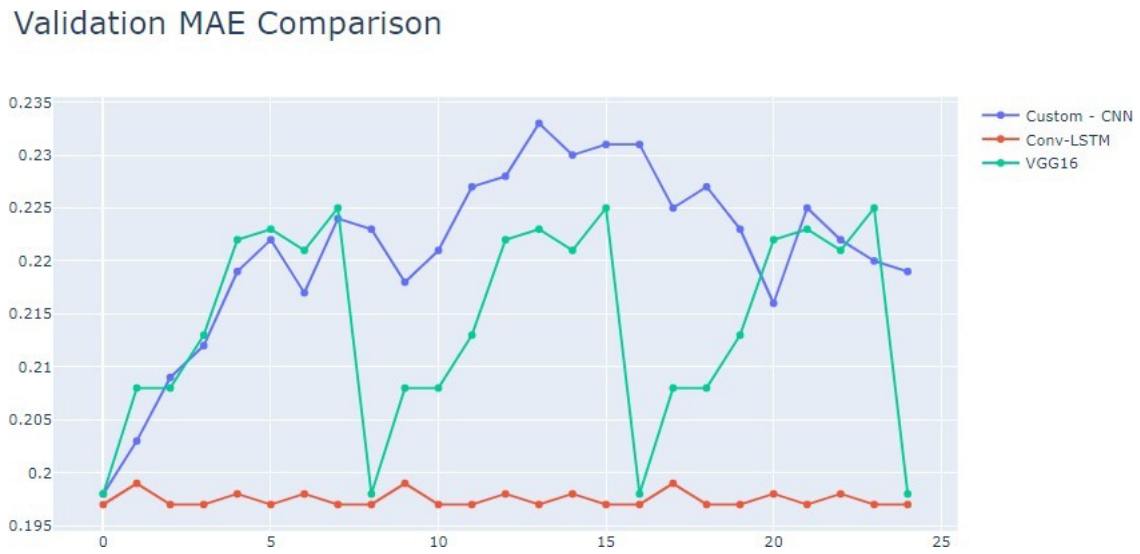

Figure 11 Validation MAE Comparison of Models

The Conv-LSTM framework showed the greatest MAE accuracy validation sets, closely followed by the Customized CNN and VGG-16 models. The findings show that the Conv-LSTM architecture's improved predicting skills were a result of the inclusion of spatial and temporal data.

6.2 Evaluation based on Mean Squared Error

When evaluating the forecasting accuracy of the computational models, the Mean Squared Error (MSE) measure takes into account the squared variations between the forecasted and real conveying values. The findings were as the Custom CNN algorithm displayed a 0.083 MSE during validation. These numbers show that, during training, the model successfully minimized the squared discrepancies between its forecasts and the actual expression values, leading to considerably fewer mistakes on the training set than on the validation set. A testing MSE of 0.063 were attained by the Conv-LSTM model. The fact that the squared error rates for predictions on

the validation set was significantly reduced implies that the model was successful in capturing the temporal connections within the video frames. The VGG-16 model showed a validation MSE of 0.064 by utilizing transfer learning. In terms of comparing the algorithms with MSE, it is clear evident that Conv-LSTM is the best performing model. The comparison of Algorithms based on validation MSE is shown in Figure 12.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

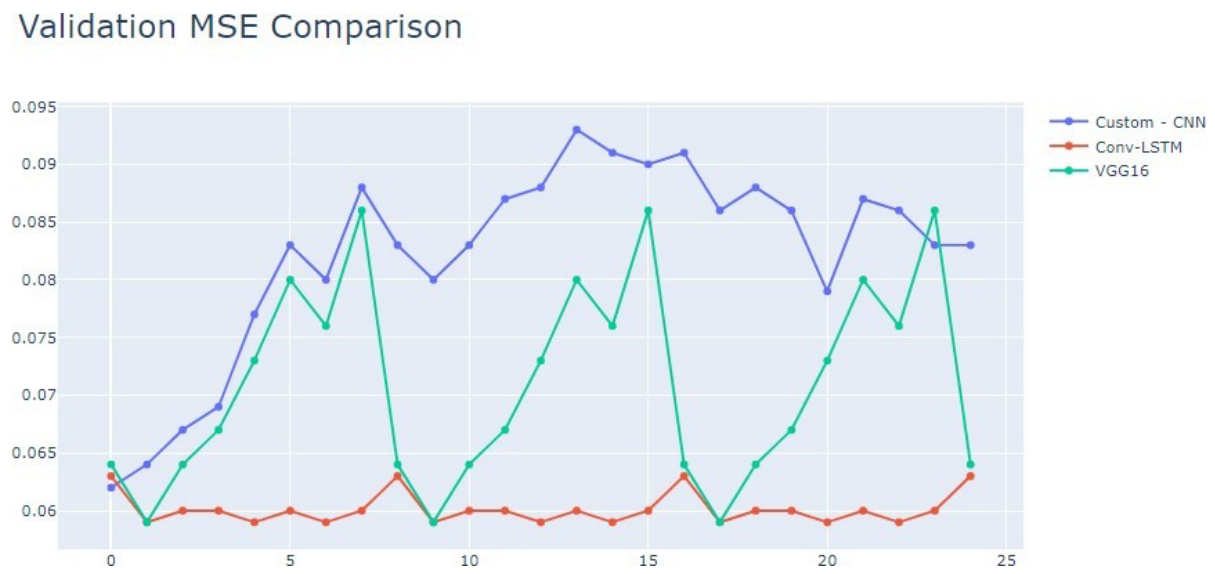

Figure 12 Validation MSE Comparison of Models

In conclusion, the Conv-LSTM model had the best MSE performance on validation sets, highlighting its skill in simulating both temporal as well as spatial dynamics. The Custom CNN and VGG-16 models also exhibited competitive performance, further highlighting the significance of leveraging various deep-learning architectures for evoked expression prediction.

6.3 Evaluation based on Loss.

The assessment utilizing loss functions provides a thorough viewpoint on the model training and optimization. The assessment centered around loss functions offers insightful information about the fluctuations of each algorithm's training and highlights their convergence towards improved prediction skills for evoked expression detection. The Custom CNN model had a validation loss of 0.082. These numbers imply that the model was successful in minimizing the difference between the expression values that were predicted and those that actually occurred during training, which helped to lower error rates for prediction on both the sets used for training and validation. A validation loss of 0.064 was shown by the Conv-LSTM model. On the other hand, a validation loss of 0.066 were shown by the VGG-16 model. The algorithm's validation loss was equivalent to that of the Conv-LSTM model, looking at precise result it has been identified that CONV-LSTM performs optimally in terms of validation loss. The results demonstrate the model's ability to modify previously taught features and enhance its performance by reducing losses. Overall, the

Conv-LSTM model had the best performance in terms of validation losses, demonstrating its capability to efficiently collect spatial and temporal data. The comparison of Algorithms based on validation Loss is shown in Figure 13.

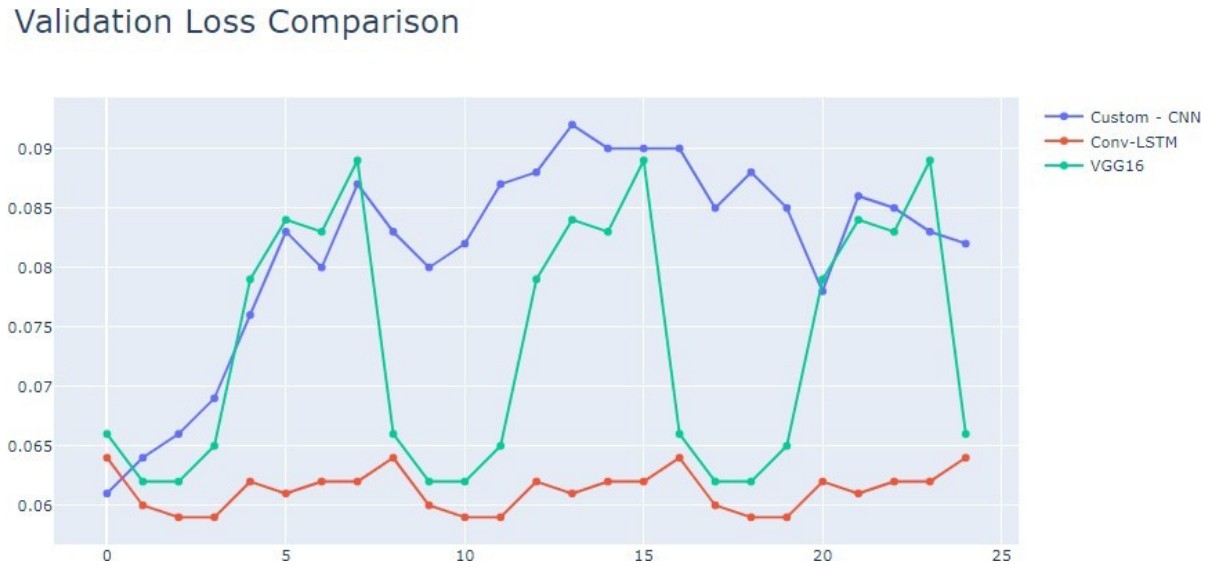

Figure 13 Validation Loss Comparison of Models

6.4 Predictions on Test Data

The evaluation of Algorithms based on the various metrics provides Implicit information about the performance of the algorithm. Therefore, it is better to make predictions using implemented models on some test data images because the visualization of the predictions on test data gets better insights about the forecasts made by the algorithms. Figures 14, 15, and 16 represent the predictions made by all the executed models on some test data (Video1, Video2, Video3). From these illustrations, it is crystal clear that the Conv-LSTM model is performing superlative against other implemented models. In order to analyze the MSE and MAE values accompanied by each model while predicting the test data on each set of videos, bar plots are plotted to represent the output values of these metrics. Figure 17 which depicts the MAE and MSE metric value of each model in bar plots, it is understandable that the metric value achieved by the Conv-LSTM model is the lowest among all other models thus making more accurate predictions than other executed models.

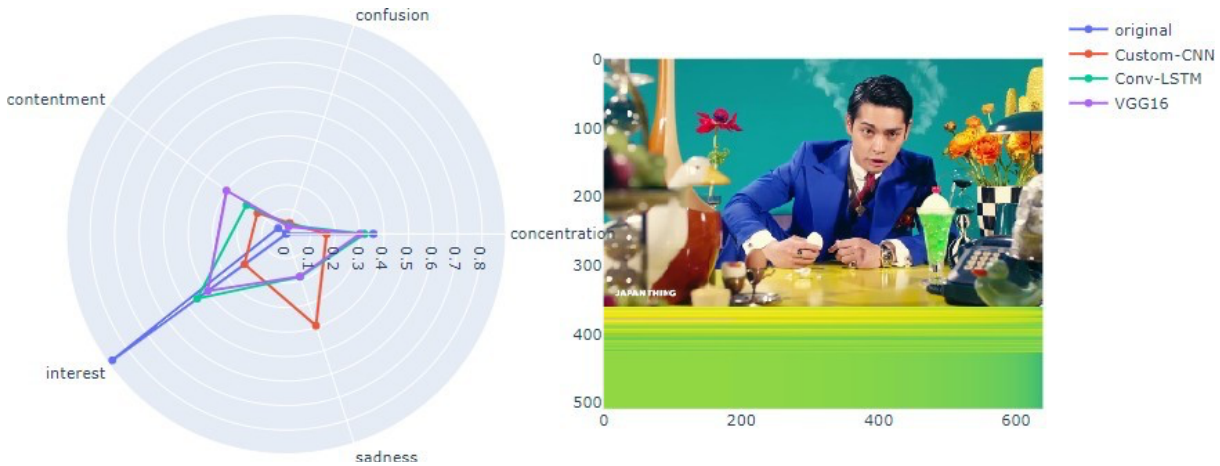

Figure 14 Predictions on Video1 By all Models

Figure 15 Predictions on Video2 By all Models

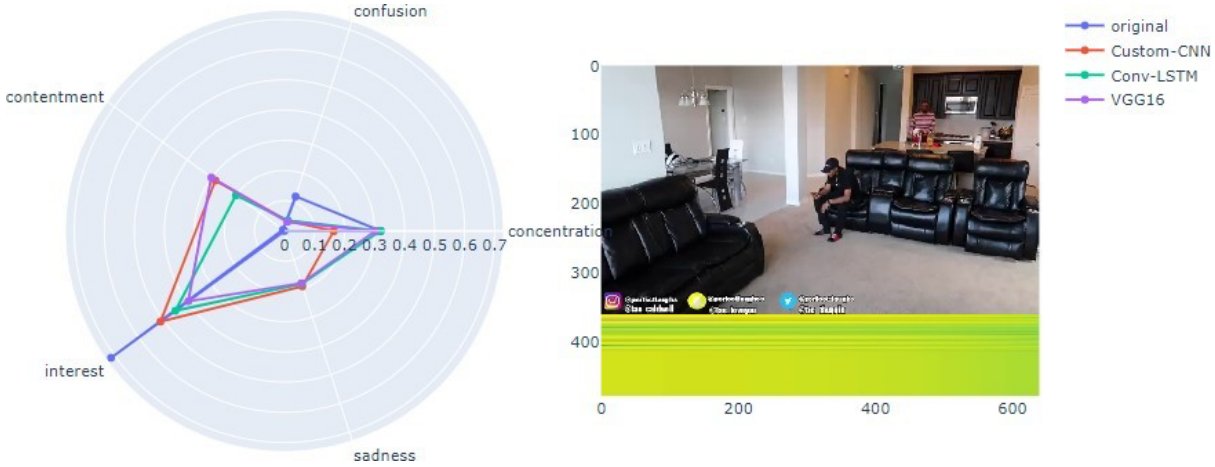

Figure 16 Predictions on Video3 By all Models

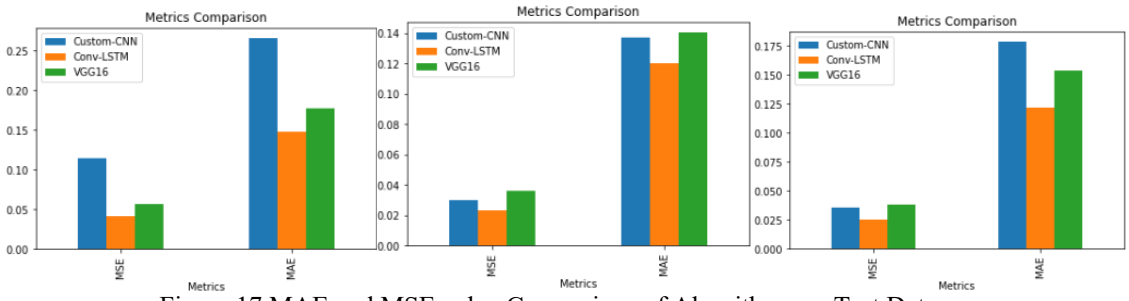

Figure 17 MAE and MSE value Comparison of Algorithms on Test Data

6.5 Discussion

The study on recognizing evoked expressions in videos set off on a thorough trip that included several stages and ended with a thorough examination of deep neural network structures. The Evoked Expressions from Videos (EEV) dataset was carefully curated in the first stage using data that was carefully acquired from AI crowd. A variety of approaches, comprising the collection of frames from films and the creation of spectral images from audio segments, were used in the ensuing data preparation step. The careful dimension normalization, pixel value scaling, and coordination of OpenCV for picture manipulation laid the foundation for further modeling efforts. The Custom CNN, VGG-16, and Conv-LSTM models were the three distinct deep-learning structures that were used in the research's modeling efforts. Every single one of these models debuted a unique combination of strategies for capturing and decoding the nuances of facial expressions within video frames. The VGG-16 algorithm model used transfer learning, which involved employing previously acquired features to improve its understanding of the nuances of expressing itself, whereas the customized CNN model probed deeply into minute details to extract special characteristics from images. Convolutional neural network (CNN) and Long Short-Term Memory (LSTM) networks were combined to create the Conv-LSTM architecture, which aimed to close the disparity across both temporal and spatial patterns. To improve their capacity for prediction, these algorithms undertook an intense training process that included 25 iterations. The Conv-LSTM model appeared as the expert at the conclusion of this operatic fall with the lowest validation MAE among its competitors with a value of 0.197. Its natural capacity to uncover the temporal and spatial patterns hidden inside video frames is a key factor in this accomplishment. Nevertheless, while displaying outstanding testing MAE values of 0.219 and 0.198, respectively, the customized CNN and VGG-16 models struggled to identify and understand the complex characteristics present in films. The culminating point of these results captures the compositional abilities of the Conv-LSTM model in merging spatial and temporal strands of information. The capacity to recognize and understand facial emotions was enhanced by this fusion, opening the door for improved evoked expression identification. Although impressive, the Custom CNN and VGG-16 models have trouble completely capturing the shifting patterns of temporal progression. Looking back, the process of dataset curation, preprocessing, modeling, and assessment reflects the subtleties and complexity involved in extracting evoked expressions from video frames, and it ultimately leads to the identification of the Conv-LSTM design as the best model in this endeavor.

Algorithms/Metrics	MAE	MSE	Validation Loss
Custom CNN	0.219	0.083	0.082
Conv-LSTM	0.197	0.063	0.064
VGG-16	0.198	0.064	0.066

7. Conclusion and Future Work

Numerous inventive applications have been discovered and developed, such as gauging public sentiment about specific items from tweets and forecasting election results using the massive amounts of text data that have just been accessible online. While language processing, as well as analysis, have advanced, motion pictures may still convey powerful emotions. Videos are now much more readily available online, and this trend will persist. Yet, they remain widely disregarded for evaluation of feelings due to the computing cost for processing and difficulties in analyzing footage from videos. Recent developments have allowed for the detection of an individual's emotional response to footage based on facial expressions and audio cues. However, gauging the emotional response from the video alone has proven difficult. In this study the subject matter of "Evoked Expression from Video," the investigation of several models based on deep learning has produced insightful findings about determining the presence of viewer expressions in video content. It is clear from the thorough analysis that the Conv-LSTM algorithm excelled at this endeavor, displaying superior results across several criteria. The Conv-LSTM architectural capacity to capture temporal as well as spatial data proved crucial in capturing the complex functioning of facial reactions, with validation MAE of 0.197, MSE of 0.063, and a validation loss of 0.064. This study emphasizes how crucial it is to take gestures' tendency to change over time into consideration, particularly when dealing with recordings. The potential for additional investigation is still quite broad despite where we proceed. Future studies can focus on improving the Conv-LSTM model's design, taking other behavioral aspects into account, or investigating hybrid strategies that combine the best features of many models. The use of cutting-edge methods like attention mechanisms may also improve the models' capacity to concentrate on pertinent face aspects. The results of this study and the effectiveness of the Conv-LSTM model ultimately serve as a crucial building block for the development of emotion detection systems, fostering a better comprehension of how to make sense of individual expressions in changing visual scenarios. Due to the limitation of computing capabilities a limited number of algorithms has been explored in thisresearch. However, in future more advanced models, such as transformer-based architectures like Vision Transformers (ViT) or BERT for video, can be used which can push the boundaries of prediction accuracy. Also, in future the technique can be developed, to personalize expression prediction based on individual user data, catering to the nuances of everyone’s way of expressing emotions.

References

Ben, X., Ren, Y., Zhang, J., Wang, S. J., Kpalma, K., Meng, W., & Liu, Y. J. (2022). Video-Based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5826–5846. https://doi.org/10.1109/TPAMI.2021.3067464

Ellis, J. G., Lin, W. S., Lin, C.-Y., & Chang, S.-F. (2019). *Predicting Evoked Emotions in Video*. https://doi.org/10.1109/ISM.2014.69

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 154–164. https://doi.org/10.18653/V1/D19-1015

GitHub. (2023). *GitHub - google-research-datasets/eev: The Evoked Expressions in Video dataset contains videos paired with the expected facial expressions over time exhibited by people reacting to the video content.* https://github.com/google-research-datasets/eev

Ho, N. H., Yang, H. J., Kim, S. H., Lee, G., & Yoo, S. B. (2021). Deep Graph Fusion Based Multimodal Evoked Expressions from Large-Scale Videos. *IEEE Access*, *9*, 127068–127080. https://doi.org/10.1109/ACCESS.2021.3107548

Huynh, V. T., Lee, G.-S., Yang, H.-J., & Kim, S.-H. (2021). *Temporal Convolution Networks with Positional Encoding for Evoked Expression Estimation*. https://arxiv.org/abs/2106.08596v1

James McDermott (2021) Hands-on transfer learning with keras and the VGG16 model, Learn Data Science - Tutorials, Books, Courses, and More. Available at: https://www.learndatasci.com/tutorials/hands-on-transfer-learning-keras/

Khan, A. R. (2022). Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges. *Information 2022, Vol. 13, Page 268*, *13*(6), 268. https://doi.org/10.3390/INFO13060268

Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, *175*, 689–694. https://doi.org/10.1016/J.PROCS.2020.07.101

Nguyen, D.-K., Ho, N.-H., Pant, S., & Yang, H.-J. (2023). *A transformer-based approach to video frame-level prediction in Affective Behaviour Analysis In-the-wild*. https://arxiv.org/abs/2303.09293v2

Phuong Thao, H. T., Balamurali, B. T., Roig, G., & Herremans, D. (2021). AttendAffectNet–Emotion Prediction of Movie Viewers Using Multimodal Fusion with Self-Attention. *Sensors 2021, Vol. 21, Page 8356*, *21*(24), 8356. https://doi.org/10.3390/S21248356

Phung. and Rhee. (2019) “A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets.” Applied Sciences, 9(21), p. 4500. DOI: 10.3390/app9214500.

Raut, N. (2019). *Facial Emotion Recognition Using Machine Learning*. https://doi.org/10.31979/etd.w5fs-s8wd

Shi, C. et al. (2022) “Learning Multiscale Temporal–Spatial–Spectral Features via a Multipath Convolutional LSTM Neural Network for Change Detection With Hyperspectral Images.” IEEE Transactions on Geoscience and Remote Sensing, 60, pp. 1–16. DOI: 10.1109/tgrs.2022.3176642

Song, Z. (2021). Facial Expression Emotion Recognition Model Integrating Philosophy and Machine Learning Theory. *Frontiers in Psychology*, *12*, 759485. https://doi.org/10.3389/FPSYG.2021.759485/BIBTEX

Sun, J. J., Liu, T., Cowen, A. S., Schroff, F., Adam, H., & Prasad, G. (2020). EEV Dataset: Predicting Expressions Evoked by Diverse Videos. *ArXiv*. https://www.researchgate.net/publication/338621225_EEV_Dataset_Predicting_Expressions_Evoked_by_Diverse_Videos

Thi, H., Thao, P., Herremans, D., & Roig, G. (2023). *Multimodal Deep Models for Predicting Affective Responses Evoked by Movies*. https://github.com/ivyha010/

Wang, C., Zhang, J., Jiang, W., & Wang, S. (2021). A Deep Multimodal Model for Predicting Affective Responses Evoked by Movies Based on Shot Segmentation. *Security and Communication Networks*, *2021*. https://doi.org/10.1155/2021/7650483