Emotional Analysis Using Hybrid Deep Learning Models

G V SHILPA
Vemana Institute of Technology
Mahayogi vemana road
INDIA

Abstract: A person's expressions of feeling are readable from their face, which is widely regarded as the most significant feature of the human body. Detecting and recognizing a person's face is more accurate and less expensive than other types of biometrics. It is possible to infer a person's intention and state based on their emotional state, thanks to a modality known as emotion. Within the realm of computer vision research, expression analysis and recognition have emerged as one of the more exciting research topics. New HCI research considers the user's emotional state to deliver a smooth interface. This study suggests a hybrid deep learning technique for emotion analysis based on face images. The suggested system is an amalgamation of VGG16 and Bidirectional LSTM techniques to classify various emotions on the face. Binary cross-entropy was used as a loss function to optimize the model. The model was taught and tested on the KDEF dataset. Another hybrid model comprising of Conv2D, Maxpooling2D, and Bidirectional LSTM models was tested on the CK+48 dataset. Both models showed efficient performance accuracy in training and testing the face emotion classifications.

Keywords: Image Processing, Convolutional Neural Network, VGG-16, Bi-LSTM, Maxpooling

Received: April 2, 2022. Revised: September 8, 2022. Accepted: October 2, 2022. Published: November 2, 2022.

1. Introduction

In current history, the mental health of university students has emerged as a societal issue that has garnered increasing attention from researchers and policymakers [1]. It is estimated that around twenty percent of college students in China suffer from some kind of mental disorder of varying severity [2]. Every year, one of the primary causes first-year college students withdraw from their studies is the presence of mental health issues and diseases [3]. In computer vision, a considerable amount of research has been done on developing systems that can discern emotions based on face pictures. The expression on a person's face may indicate a variety of things, including enjoyment, business, health care, and education, to understand that person's state of mind [4]. The problem of recognizing emotions from face photographs has become more difficult due to the poor quality of the photos and the variety of backdrops. In recent years, there has been an increase in the availability of various electronic devices in our market. People's lives have caused them to spend more time engaging in activities such as online shopping, gaming, and social networking. Nevertheless, the vast majority of current it has been shown that human-computer interaction (HCI) systems are not very good at understanding and comprehend emotional information, as well as a lack of emotional intelligence. This is because they cannot recognize human emotional states and information that can be used to make decisions about their emotional states.

Thus, it becomes essential for advanced intelligent HCI to find a way to fix the lack of rapport that exists between people and robots. Any HCI solution that completely disregards the user's emotional states would not react appropriately to such conditions if they were present. To address this and to solve this challenge in HCI, we must provide robots with the capability to deduce and comprehend the expressive states of human beings. Hence, a smart Human-Computer Interaction requires a dependable, precise, flexible system and a comprehensive mechanism for identifying emotions. More and more academics in artificial intelligence are working toward the ultimate aim of giving machines feelings. Artificial intelligence (AI) researchers have been studying affective computing

in general and emotion identification in particular, which positions them as an expert in these fields that are developing and are potentially fruitful study fields.

Lately, emotion-aware intellectual systems have been utilized in several fields, including e-health, e-learning, recommender systems, and smart homes, among others. Intelligent conversational systems, connected homes, and connected cities (e.g., chatbots) [5-8]. In numerous smart systems, such as online gaming, neuro-marketing (the evaluation of customers' feedback), and psychological well-being observing, the application of computer-based automated emotion identification has a significant amount of potential application. In order to create effective intervention strategies for mental health conditions, researchers are now working to identify human emotions in a precise manner considering the significance of maintaining good mental health in today's communities. For illustration, in a medical system equipped with a module for recognizing emotions, patients' mental and physical conditions may be constantly measured, and the necessary therapy can be administered per the findings. The objective of emotion recognition and detection in the field of human-computer interaction (HCI) is to create and deploy intelligent systems with optimal HCI that can adapt to users' emotional states.

Approximately twenty or thirty years ago, these open-ended topics were the exclusive purview of science fiction. Despite this, the automatic recognition and detection of human emotions has emerged as a prominent research issue during the past two decades, intending to make human-computer interaction (HCI) as natural as human interactions.

2. Literature Review

The extensive body of research on emotional recognition and analysis covers the substantial ground. For example, Mostafa *et al.* (2018) attempted to detect facial emotions in videos and distinguish between the following five features: sadness, amusement, fear, disgust, and anger [9]. The authors use various Machine Learning models to accomplish this, like RF, SVM, kNN, and Recurrent Neural Networks (i.e., RNN). Assuming the Euclidean metric, this study classifies emotions based on the state of the eyes, the mouth, and other facial features. Amusement, for instance, can be indicated by rapid eye blinking, pushed-up cheeks, crows-feet, wrinkles near the edges of the eyes, etc. The LSTM classifier performs best and can distinguish between two emotions with a minimum accuracy of 61% and a maximum accuracy of 82%.

Alakus *et al.* performed emotional recognition based on electroencephalography (EEG) signals and used the GAMEEMO dataset to predict emotions on the basis of positivity and negativity [10]. After collecting the GAMEEMO dataset's EEG signals, the study calculates the spectral entropy values for each of them. It uses Bi-LSTM to classify these values and the final model achieves an accuracy score of 76.91%, a specificity score of 76.89%, and a sensitivity score of 76.93%.

Ranganathan *et al.* (2016) present the emoFBVP database and develop a 4-Deep Belief Network model. It was shown that convolutional DBNs can be effectively employed to recognize subtle emotions. It offered a Low-intensity emotion recognition using a convolutional DBN model and presented an accuracy of 83.18% [11]. Kumar *et al.* (2017) proposed a model based on image pre-processing by recognizing faces with the Viola-Jones technique, clipping the most prominent face, and scaling to 48x48. This image is then input to the Deep 10-layer CNN model, and the error gradients for all the weights are computed through Back Propagation. The model had an accuracy of more than 90% [12].

Then came the AlexNet-based model. Huang *et al.* (2019) proposed this model was tested on a Real face dataset created by querying Yahoo, Google, and Bing with emotion-related tags in multiple languages. These datasets are combined to generate a hybrid model-training dataset. In the best training approach, the maximum accuracy received was 87.79%, and the minimum accuracy was 74.19% [13]. Pranav E *et al.* (2020) proposed model uses two convolutional layers that have a dropout after each. Keras is employed for building and fitting the model. The proposed model also uses an Adam optimizer to reduce the loss function [14]. The manually collected image dataset used in this study had five emotions – surprised, angry, sad, happy, and neutral. In order to get a more accurate categorization of feelings, Faten *et al.* (2019) investigated the effect that pre-processing the input had on the neural network before it was trained [15].

Akhand *et al.* (2021) suggest deep CNN for FER across multiple accessible databases [16]. Following the extraction of the facial characteristics from the datasets, the pictures they used for demonstration decreased in size to 48 pixels on each side. After that, they carried out the procedure of augmentation of data. The architecture that was utilized

consists of two layers of convolution pooling, followed by the addition of two modules of the inception style, each of which has convolutional layers of sizes 1x1, 3x3, and 5x5. They demonstrate the potential to employ a technique known as the network-in-network, which enhances local performance owing to the convolution layers applied locally. Kaur *et al.* (2021) proposed a model utilizing CNN to evaluate face expressions and a time-distributed convolutional neural network with augmented data and early stopping techniques for accuracy enhancement for evaluating audio features [17]. It had an accuracy of 73% for facial expressions and 90.91% for audio features.

The proposed model in the study presented by Do *et al.* consists of the face feature model, the LSTM Module, and the STAT Module [18]. The face feature model uses VGGFace2 ResNet50. It is trained on AffecNet and RAF-DB datasets and outputs the emotion probability scores. These probability scores are then fed to the LSTM Module and the STAT module. The proposed model can distinguish between neutral, angry, disgusted, fearful, happy, sad, and surprised emotions. Liu *et al.* (2020) proposed a hybrid CNN-BiLSTM model for the emotional analysis of the text. It is divided into three parts, namely the word vectorization module, the feature selection, and extraction module, and finally, the classifier [19], yielding an accuracy of 94.2%.

3. Deep Learning Models for Case Study

Convolutional Neural Networks (CNN)

Deep learning methods are based on neural networks, a branch of machine learning but are nonetheless essential to the process [20]. They are mainly composed of node levels, each of which has an input layer, single or multiple hidden layers, and an outcome level. Every node is connected to every other node and has a weight and an associated threshold [21]. Suppose the outcome of any one individual node is greater than the value that has been determined to be the threshold for activation. In that case, that node will become active and begin passing data to the subsequent model level.

CNN models are distinguished by their more excellent performance when given inputs involving images than other types of neural networks [22-24]. They are composed of three primary sorts of strata, which are as follows:

3.1 Layer of Convolutional Data

The convolutional layer is where most CNN computing occurs [25]. This input will have height, width, and depth, which match RGB in an image, assuming the information is a 3D color pixel matrix, A feature finder, also called a kernel [26], checks the image's receptive fields for the feature. Convolution describes this process.

The feature finder is a 2-D array of image weights. Filter size is commonly a 3x3 matrix, which determines the reception field. A dot product is computed between the input pixels and the kernel. The output array receives the dot product. The filter then moves by a stride until the kernel sweeps the entire image [27]. The ultimate output is a feature map, activation map, or convolved feature [28].

3.2 Pooling layer

Dimensionality reduction is achieved by pooling layers [29, 30], which is often referred to as down sampling [31-33]. This technique cuts down on the number of parameters that are used in the input. The pooling operation works very similarly to the convolutional layer in that it applies a filter to the entire input [34]; however, unlike the convolutional layer, this kernel have no weights associated with it. In its place, the kernel uses an aggregation function, which applies to the data contained inside the reception field to populate the output array. Pooling can be broken down into two primary categories: Max-pooling and Average-pooling [34, 35].

3.3 Layer that is Fully Connected (FC)

The full-connected layer's name accurately depicts its function. There is no direct connection between the input pixel values and the output units in partially connected layers. A node at each output layer is connected to another node at the preceding layer via a direct connection in the fully-connected layer [36]. This layer is responsible for performing the classification process using the characteristics that were retrieved from the layers that came before it using their respective filters [37]. FC layers typically use a softmax activation function [38, 39] to accurately categorize inputs, which results in a probability ranging from 0 to 1; this is in contrast to the tendency of convolutional and pooling layers to make use of ReLu functions.

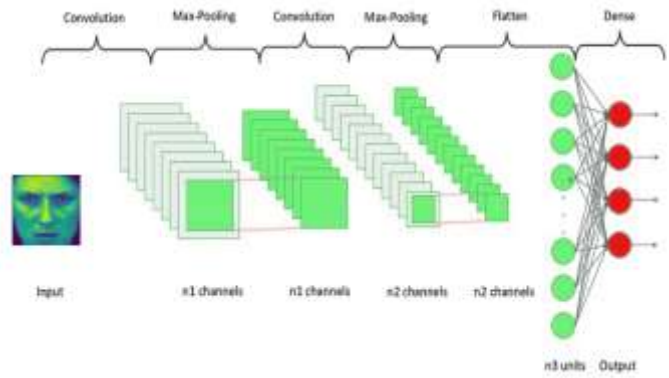


Fig. 1. A pictorial representation of CNN architecture flow-diagram (*Source :* [40])

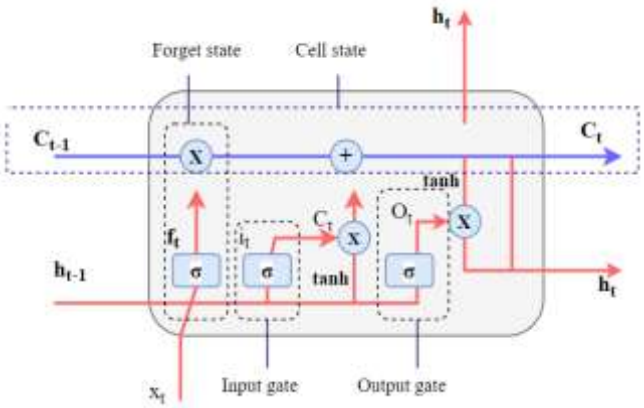


Fig. 2. A simple flow diagram of the LSTM network

Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) neural network is a specialized form of the deep learning model (RNN) [41] that can manage protracted interconnections, particularly in issues involving sequence prediction [42]. LSTM is equipped with feedback connections, which means it can process the entirety of the data sequence, except for single data points like images. The structure of a recurrent neural network always takes the form of a series of modules of the neural network that are repeated. Even though LSTMs have this chain-like form, the structure of the repeating module is quite different [43]. There are numerous layers of neural networks rather than just one, and these layers distinctly interact with one another.

An LSTM model's most important function is played by a memory module that stays in the same state for an extended period and is referred to as a 'cell state' [44]. The cell state is represented by the horizontal line traversing the top of the diagram, which may be found below in Fig. 2. It is possible to imagine it as a conveyor belt across which information passes without being altered in any way. Gates add or remove information from LSTM cells [45]. These gates allow information to enter and leave cells. It uses point-wise multiplication (X), point-wise addition (+) as well as a sigmoid neural net layer (σ) to aid the procedure. The sigmoid layer outputs integers between zero and one, with zero indicating that nothing ought to be allowed to pass through. Here, one indicates that everything ought to be permitted to go through.

LSTM can be implemented on several types of time series problems such as Univariate, Multivariate etc. Further various approaches of LSTM models are there to solve univariate cases. They are-Data Preparation [46], Vanilla LSTM [47], Stacked LSTM [48], Bidirectional LSTM [49], CNN LSTM [50], and ConvLSTM [51].

4. Datasets and Data Pre-Processing and Augmentation

We used different models' datasets in this study to compare the outcomes. The datasets for the models' implementation are discussed below in detail.

Jaffe dataset

The Japanese Female Facial Expression (JAFFE) database was compiled using photos of 213 varied facial gestures contributed by 10 different Japanese female participants. This data was made available to the public. Every participant experienced the six fundamental feelings, plus neutral, in a series of expressions, with 30 representing anger, 29 representing disgust, 33 representing fear, 30 representing happy, 31 representing sadness, 30 representing surprises, and 30 representing neutral. The resolution of the image is presented in grayscale. Every single photograph of a person's face was captured in identical lighting and without any occlusions, such as hair or glasses. These settings were carefully regulated. The resolution of the original photograph is 256 pixels on each side, and all of the facial expressions can be seen in frontal view.

CK+48 Dataset

The Extended Cohn-Kanade (CK+) dataset contains a total of 593 video streams collected from 123 separate participants. These subjects range in age from 18 to 50 years old, and they come from a diversity of genders and ethnic backgrounds. Each video displays a face transition from a moderate expression to a specified maximal expression. The recordings were recorded at a frame rate of thirty frames per second (FPS), and their quality is either 640 by 490 or 640 by 480 pixels. 327 of these films have been categorised into one of these seven categories of emotional expression: fury, disdain, disgust, fear, pleasure, grief, and astonishment are some of the emotions that might be experienced. The CK+ database is utilised in the vast majority of facial expression classification strategies, and it is commonly recognised as the facial gestures categorization dataset that is used the most extensively and is produced by a laboratory that controls its conditions.

KDEF Dataset

The Karolinska Directed Emotional Faces (KDEF) database is another collection of data that is accessible to the public and contains a collection of 4900 photographs of face expressions. It consists of seventy people, each presenting seven distinct facial expressions, and each expression being photographed (twice) from five different perspectives. Before the model is applied on the KDEF dataset, it is loaded for farther pre-processing. There are 8 classes in the KDEF dataset. They are, 'anger', 'contempt', 'disgust', 'feat', 'happy', 'neutral', 'sadness', 'surprise'. These classes are put into an array.

CV2 Cascade Classifier

After the array step, the 'haarcascade' classifier is applied on the dataset [52]. The 'haarcascade' algorithm is one of the more well-known facial recognition algorithms [53-55]. It is a fast algorithm that also provides a high level of accuracy, and it has a lower computational cost. A dark region and a light region make up the components of a haar-like feature [56]. It does this by calculating the difference between the total strengths of the darker sections and the total strengths of the brighter sections, and using that information to construct a single number. This is done in order to retrieve valuable aspects that are required for the identification of an object. This classifier detects face and eye in the image data.

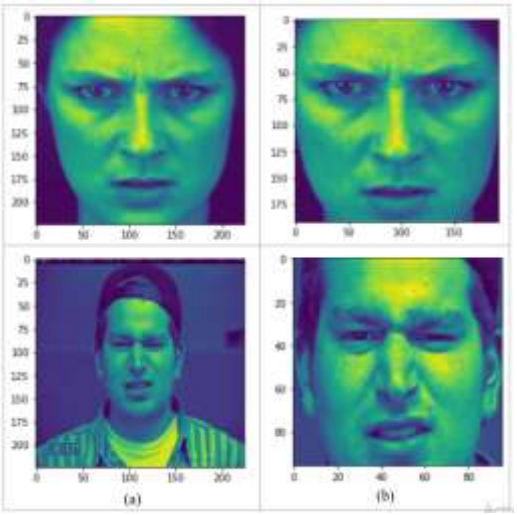


Fig. 3. (a) Before Applying Cascade Classifier
(b) After Applying Cascade Classifier.

Data Augmentation

The performance of most deep learning models is contingent on the training data's quality, amount, and relevance. Nevertheless, a lack of data is among the most typical issues when deploying deep learning in an industrial setting. This is due to the fact that the collection of such data can, in many instances, be both expensive and time-consuming. This is the reason data augmentation is applied to the dataset [57-59].

Data augmentation is a collection of methods used to artificially generate samples from pre-existing data to enhance the total amount of data [60]. This may involve making relatively minor data adjustments or using deep learning models to produce new data points [61]. Some of the advantages of augmenting data are as follows:

1. It enhances model prediction efficiency.
2. It incorporates even more data for training into the models.
3. It reduces the amount of missing data in order to build more accurate models.
4. It lowers the likelihood of data overfitting. This statistical error occurs when a function correlates too close to a predetermined group of

data points, increasing the amount of variability present in the data.

5. It helps improve the models' ability to generalize, which assists in tackling class imbalance problems in classification.

6. It brings down the costs of gathering and classifying data making it possible to anticipate unusual events.

7. It assists in preventing data privacy concerns.

We set the degree range for random rotations to be 20 for this dataset before implementing it in the hybrid model in the research. The field for random zoom was set at 0.05, and the shear intensity range was fixed at the same. The image data was allowed to flip horizontally and vertically. Rescale factor was set at 1/255.0, keeping the data at that scale after applying all the other transformations. 10% of the image data were reserved for model validation after training. Points outside the boundaries of the input data were filled using the nearest mode. The brightness range for the images was given as a tuple value (two floats). This range was from 0.3 to 1.

5. Model Architecture

In this study, Tensorflow was used to implement and configure the models. The data pre-processing part was on Keras. Keras layers were utilized to design CNN models. For the evaluation of the performance of the model, we applied Sklearn metrics. GPU resources were used to execute the modules.

Conv2D + Maxpooling2D + Bi-LSTM MODEL

A deep convolutional bidirectional long short-term memory (Bi-LSTM) fusion network that is capable of making use of both spatial and temporal input for the recognition of facial expressions (FER). The deep temporal network (DTN) looks at changes in short-term expression and takes as input two frames that are immediately following one another. When presented with a series of images that represent a certain emotion class, the deep spatial network (DSN) is able to learn the detailed attributes of each frame. When compared to the DTN, the DSN possesses both a greater depth and a larger architecture. After that, a Bi-LSTM network is utilised to discover data correlations for the purpose of enhancing the capability to learn information that is spatially and temporally discriminative. After that, the softmax classifier is applied to the problem in order to place the given sequence into one of the primary expression classes. In order to acquire knowledge regarding the properties of the spatial appearance, the DSN architecture operates on individual frames by means of a number of expressions. DTN receives as input a pair of frames that have occurred successively in a series of expressions. These frames have been viewed in sequence. The short-term temporal variations in appearance are recorded in an explicit manner by this type of input. A shallower than usual depth is used in the DTN model in particular. The temporal information that occurs between two consecutive frames is the primary focus of DTN, which most certainly does not require a wider network than DSN. [62].

In this case, the input data was fed into the Conv2D model. Then the data passes through one batch normalization layer and an activation layer. This block of process repeats twice. After that comes a max-pooling layer. Again the previous block runs twice. Then a two-max-pooling layer follows another block run. Then, the Bi-LSTM model is applied to this reshaped dataset.

Novel VGG 16 + Bi-LSTM Hybrid Model

Emotion analysis is an essential subject of study within the science of natural language processing. Bi-LSTM-CNN solves the challenge of major corporate categorization. The Bi-LSTM model obtains the representation of two directions, which is then mixed using a convolution neural network to get a new expression. In addition to addressing the issues of gradient disappearance and gradient explosion, Bi LSTM considers all of the information currently available in the context.

The VGG16 CNN comprises five convolutional blocks, each including two or three convolutional layers for feature extraction. In each block, the convolutional layers have the same number of filters; after each max-pooling layer, the number of filters doubles. All convolutional layers contain identical 3 x 3 receptive fields, which increases the nonlinearity of the feature extraction module. In

this network, the max-pooling layers are used to minimize the size of feature maps, hence halving the width and height of the feature map. The primary goal of such a design is to extract a large number of different characteristics while using as few computing resources as possible. Following the feature extraction module comes the classification module. Due to form limits, the CNN feature extractor output is not used directly as an input to the Bi-LSTM layers. Hence, a reshape layer is utilised to connect the CNN module to the Bi-LSTM layers [63]. Then the Flatten layer and dense layers are applied to get the emotion classifications as output.

The novel hybrid model was tested on the KDEF dataset. VGG16 was set as the base model where 'imagenet' was utilised as weights.

Novel VGG 16 + Bi-LSTM Hybrid Model for KDEF Datasets

Input: KDEF augmented training and testing dataset
Output: Classification of image data on the basis of emotion

Step 1: loading of image data

Step 2: Data augmentation

Step 3: Train data generator for producing segmented image data batches from training, testing, and validation data.

Step 4: Employ VGG16+ Bi-LSTM
Include the three fully-connected layers as false

Step 5: Convert 3-D feature maps to 1-D using flatten layer

Step 6: Stick two fully connected layers on top of it

Step 7: Sigmoid activation of model for binary classification

Step 8: Binary cross-entropy as a loss function to optimize the model

$$Loss = -\frac{1}{output}\sum_{i=1}^{output\ size} y_i . \log \hat{y}_i + (1 - y_i) . \log(1 - \hat{y}_i)$$

Step 9: Monitor validation loss and restore the best weights

Step 10: Save the best weights after training

Step 11. Visualize loss by plotting graphs

Step 12: Prediction of test-image data

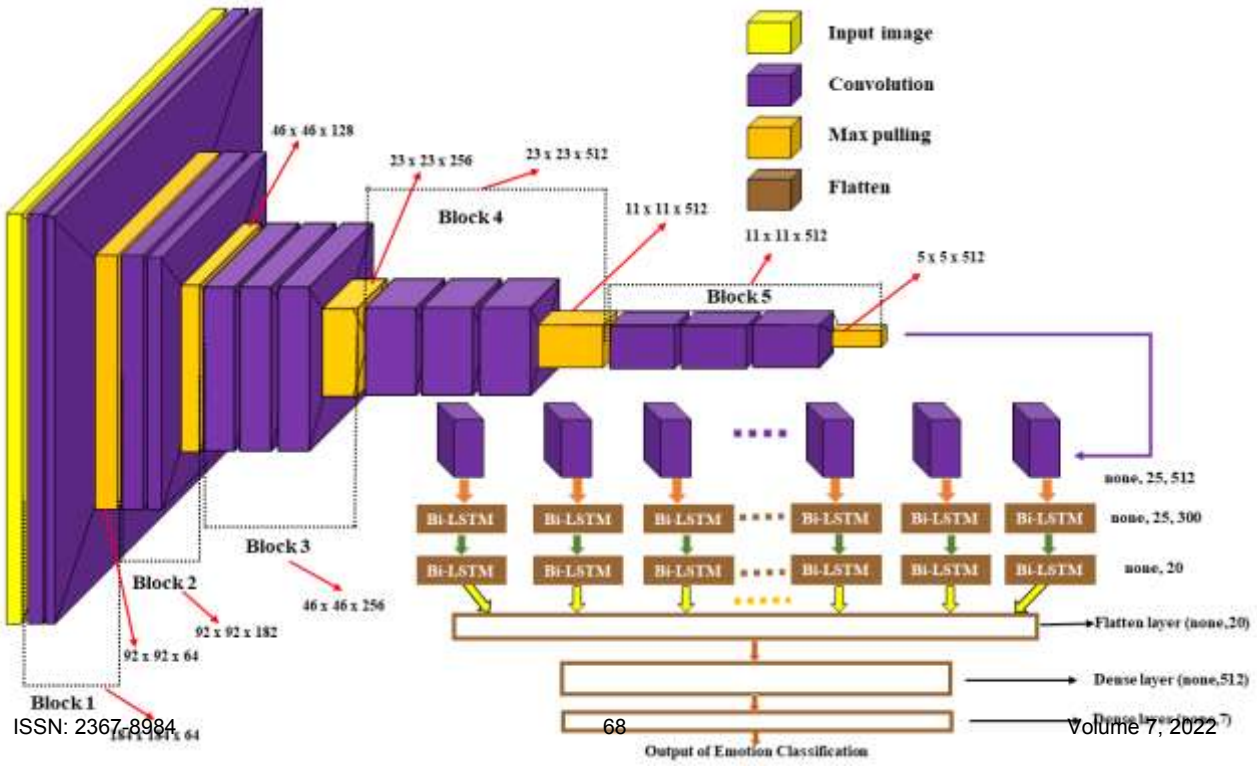
Fig. 4. The simplified algorithm of the proposed novel hybrid model (VGG 16 + Bi-LSTM)

In the above Fig. 4 the simplified algorithm of the proposed hybrid model is presented.

Fig. 5. Construction of VGG 16 + Bi-LSTM Hybrid Model

6. Results and Discussion

In this study we tested different datasets on different models to compare the performance accuracy and model efficiency.



Results of Conv2D + Maxpooling2D + Bi-LSTM MODEL for Jaffe Dataset

The Conv2D added with Maxpooling and Bidirectional- LSTM model worked well with Jaffe dataset. It worked well for the classes like 'surprise', 'fear', and 'angry'. The f1-scores for these classes came out as commendable but it needs more development on pre-processing and the model optimization. In the figure below (Fig. 6) the plots comparing losses and accuracy for training and testing datasets are shown.

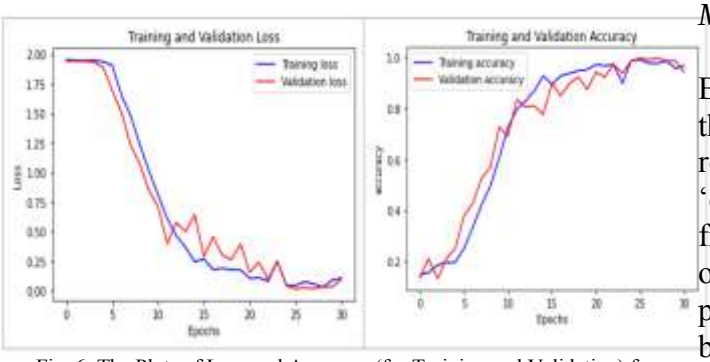


Fig. 6. The Plots of Loss and Accuracy (for Training and Validation) for Conv2D + Maxpooling2D + Bi-LSTM MODEL when deployed on Jaffe dataset

Results of Conv2D + Maxpooling2D + Bi-LSTM MODEL for CK+48 Dataset

In comparison with the Jeffe dataset, CK+48 dataset worked far better with the Conv2D + Maxpooling2D + Bi-LSTM hybrid model. The prediction for emotion recognition using this model gave an accuracy as high as 93%. It recognized 166 different image emotions correctly. The classes which were most accurately predicted using this model were- 'disgust', 'anger', ''happy', 'surprise' etc. Other classes gave high f1 score, precision, and recall values making the overall performance accuracy really high. In the Fig. 7 the plots represent the loss and accuracy comparison in cases of training and testing dataset when the model is deployed to the CK+48 dataset.

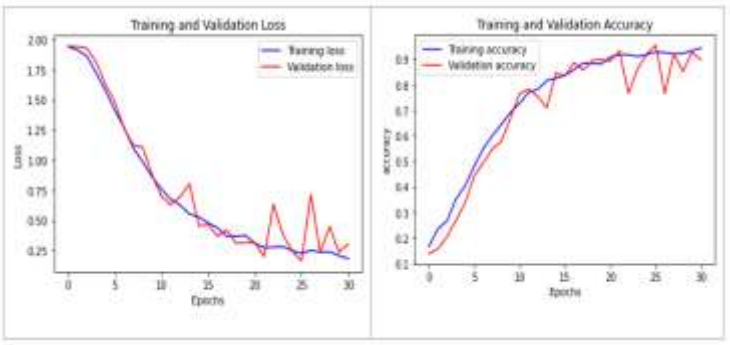


Fig. 7. The Plots of Loss and Accuracy (for Training and Validation) for Conv2D + Maxpooling2D + Bi-LSTM MODEL when deployed on CK+48 dataset

Results of Novel VGG 16 + Bi-LSTM Hybrid Model

The hybrid model comprised of VGG16 and Bidirectional LSTM techniques, was employed to the KDEF dataset. The model gave a satisfactory result. The predictions for 'surprise', 'anger', 'disgust' and 'neutral' were the most accurate ones from this model. The model needs more optimization in future as well as modified data pre-processing and data augmentation to get further better accuracy. In below figures (Fig. 8 and 9) the plots are showing comparison of training and testing datasets for both of loss and accuracy. Fig. 9 represents various classification reports from the model output.

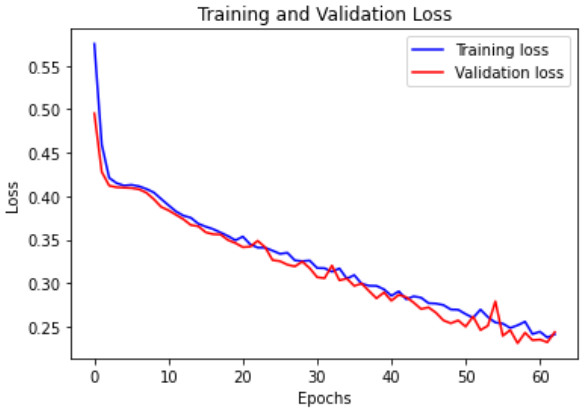


Fig. 8. The plot of Training and Validation loss for proposed model

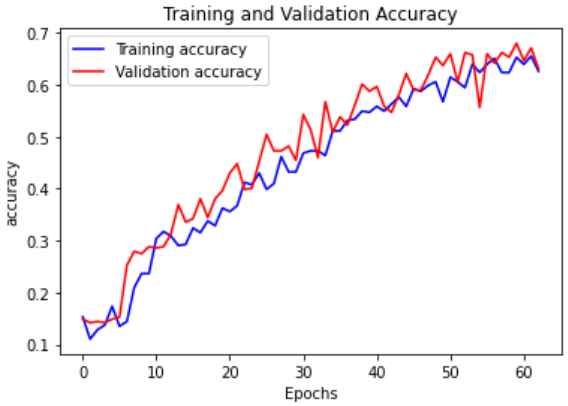


Fig. 9. The plot of Training and Validation Accuracy for proposed model

Classification Report

	precision	recall	f1-score	support
anger	0.62	0.50	0.56	10
disgust	1.00	0.77	0.87	13
fear	0.80	0.33	0.47	12
happy	0.73	1.00	0.85	11
neutral	0.71	1.00	0.83	12
sadness	0.57	0.73	0.64	11
surprise	0.92	0.92	0.92	12
accuracy			0.75	81
macro avg	0.76	0.75	0.73	81
weighted avg	0.77	0.75	0.74	81

Fig. 10. The Classification Report of Prediction for Test Images from Proposed Model (VGG16 and Bi-LSTM hybrid model)

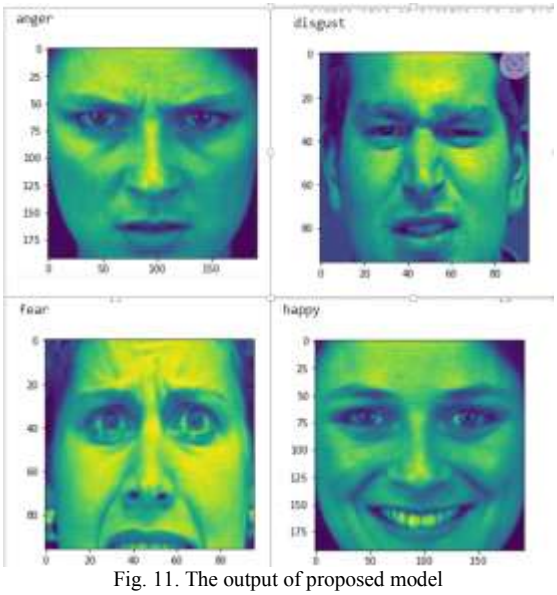


Fig. 11. The output of proposed model

The output images are shown with their classification names in the Fig. 11. This indicates that the hybrid model works well with extreme emotions expressed on face.

7. Conclusion

In this paper, various different studies aiming to detect emotion recognition using machine learning were discussed and their analysis as well as comparisons were provided. In their own limits, many of these studies demonstrated a very efficient and accurate means of doing so. The Conv2D + Maxpooling2D + Bi-LSTM hybrid model showed a high accuracy of 93% making it an efficient case study for face expression recognition. Although the VGG 16 + Bi-LSTM hybrid model shows promising results when employed to the KDEF dataset, the model needs more improvement in future to get better accuracy and f1 score.

Thus, this research drives us to conclude that Facial Expression Recognition (FER) is among the most essential means of delivering information about an individual's emotional state; nevertheless, on the whole, only seven fundamental emotions and neutral can only be learned by these methods, therefore it is always restricted. It is in direct opposition to what is experienced in day-to-day life, which is characterized by more nuanced feelings. Because of this, researchers in the future will be driven to strive toward building larger databases and developing sophisticated deep learning models to detect all primary and secondary feelings.

References

[1] P. J. C. Sahu, "Closure of universities due to coronavirus disease 2019 (COVID-19): impact on education and mental health of students and academic staff," vol. 12, no. 4, 2020.

[2] X. Liu, S. Ping, W. J. I. j. o. e. r. Gao, and p. health, "Changes in undergraduate students' psychological well-being as they experience university life," vol. 16, no. 16, p. 2864, 2019.

[3] A. Duffy *et al.*, "Predictors of mental health and academic outcomes in first-year university students: Identifying prevention and early-intervention targets," vol. 6, no. 3, 2020.

[4] I. Camerlink, E. Coulange, M. Farish, E. M. Baxter, and S. P. J. S. r. Turner, "Facial expression as a potential measure of both intent and emotion," vol. 8, no. 1, pp. 1-9, 2018.

[5] S. Munoz, O. Araque, J. F. Sánchez-Rada, and C. A. J. S. Iglesias, "An emotion aware task automation

architecture based on semantic technologies for smart offices," vol. 18, no. 5, p. 1499, 2018.

[6] M. W. Moreira, J. J. Rodrigues, N. Kumar, K. Saleem, and I. V. J. I. F. Illin, "Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems," vol. 47, pp. 23-31, 2019.

[7] P.-S. Chiu, J.-W. Chang, M.-C. Lee, C.-H. Chen, and D.-S. J. I. A. Lee, "Enabling intelligent environment by the design of emotionally aware virtual assistant: A case of smart campus," vol. 8, pp. 62032-62041, 2020.

[8] D. Ayata, Y. Yaslan, M. E. J. J. o. M. Kamasak, and B. Engineering, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," vol. 40, no. 2, pp. 149-157, 2020.

[9] A. Mostafa, M. I. Khalil, and H. Abbas, "Emotion recognition by facial features using recurrent neural networks," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 2018, pp. 417-422: IEEE.

[10] T. Alakus and I. J. E. L. Turkoglu, "Emotion recognition with deep learning using GAMEEMO data set," vol. 56, no. 25, pp. 1364-1367, 2020.

[11] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-9: IEEE.

[12] G. R. Kumar, R. K. Kumar, and G. Sanyal, "Facial emotion analysis using deep convolution neural network," in *2017 International Conference on Signal Processing and Communication (ICSPC)*, 2017, pp. 369-374: IEEE.

[13] C.-C. Huang, Y.-L. Wu, and C.-Y. Tang, "Human Face Sentiment Classification Using Synthetic Sentiment Images with Deep Convolutional Neural Networks," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2019, pp. 1-5: IEEE.

[14] E. Pranav, S. Kamal, C. S. Chandran, and M. Supriya, "Facial emotion recognition using deep convolutional neural network," in *2020 6th International conference on advanced computing and communication Systems (ICACCS)*, 2020, pp. 317-320: IEEE.

[15] F. Khemakhem and H. Ltifi, "Facial expression recognition using convolution neural network enhancing with pre-processing stages," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 2019, pp. 1-7: IEEE.

[16] P. Navdeep, N. Sharma, and M. Arora, "Facial Emotions Recognition System using Hybrid Transfer Learning Models and Optimization Techniques," in *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2022, vol. 2, pp. 182-187: IEEE.

[17] S. Kaur and N. Kulkarni, "A Deep Learning Technique for Emotion Recognition Using Face and Voice Features," in *2021 IEEE Pune Section International Conference (PuneCon)*, 2021, pp. 1-6: IEEE.

[18] N.-T. Do, T.-T. Nguyen-Quynh, and S.-H. Kim, "Affective expression analysis in-the-wild using multi-task temporal statistical deep learning model," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 624-628: IEEE.

[19] Z.-x. Liu, D.-g. Zhang, G.-z. Luo, M. Lian, and B. J. C. C. Liu, "A new method of emotional analysis based on CNN–BiLSTM hybrid neural network," vol. 23, no. 4, pp. 2901-2913, 2020.

[20] Y. LeCun, Y. Bengio, and G. J. n. Hinton, "Deep learning," vol. 521, no. 7553, pp. 436-444, 2015.

[21] T. Geng, T. Wang, A. Sanaullah, C. Yang, R. Patel, and M. Herbordt, "A framework for acceleration of CNN training on deeply-pipelined FPGA clusters with work and weight load balancing," in *2018 28th international conference on field programmable logic and applications (FPL)*, 2018, pp. 394-3944: IEEE.

[22] K. O'Shea and R. J. a. p. a. Nash, "An introduction to convolutional neural networks," 2015.

[23] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*, 2017, pp. 1-6: Ieee.

[24] S. Hijazi, R. Kumar, and C. J. C. D. S. I. S. J. Rowen, CA, USA, "Using convolutional neural networks for image recognition," vol. 9, 2015.

[25] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5-10.

[26] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2921-2930.

[27] J. Hannink *et al.*, "Mobile stride length estimation with deep convolutional neural networks," vol. 22, no. 2, pp. 354-362, 2017.

[28] X. Li *et al.*, "Delta: Deep learning transfer using feature map with attention for convolutional networks," 2019.

[29] B. Zhao, X. Dong, Y. Guo, X. Jia, and Y. J. N. P. L. Huang, "PCA dimensionality reduction method for image classification," vol. 54, no. 1, pp. 347-368, 2022.

[30] H. Gholamalinezhad and H. J. a. p. a. Khosravi, "Pooling methods in deep neural networks, a review," 2020.

[31] Y. Zhang, D. Zhao, J. Zhang, R. Xiong, and W. J. I. T. o. I. P. Gao, "Interpolation-dependent image downsampling," vol. 20, no. 11, pp. 3291-3296, 2011.

[32] D.-X. J. N. N. Zhou, "Theory of deep convolutional neural networks: Downsampling," vol. 124, pp. 319-327, 2020.

[33] D. Marin *et al.*, "Efficient segmentation: Learning downsampling near semantic boundaries," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2131-2141.

[34] Z. Song *et al.*, "A sparsity-based stochastic pooling mechanism for deep convolutional neural networks," vol. 105, pp. 340-345, 2018.

[35] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *International conference on rough sets and knowledge technology*, 2014, pp. 364-375: Springer.

[36] S. S. Basha, S. R. Dubey, V. Pulabaigari, and S. J. N. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," vol. 378, pp. 112-119, 2020.

[37] Q. Xu, M. Zhang, Z. Gu, and G. J. N. Pan, "Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs," vol. 328, pp. 69-74, 2019.

[38] I. Kouretas and V. Paliouras, "Simplified hardware implementation of the softmax activation function," in *2019 8th international conference on modern circuits and systems technologies (MOCAST)*, 2019, pp. 1-4: IEEE.

[39] Y. Li *et al.*, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10991-11000.

[40] M. T. García-Ordás, J. A. Benítez-Andrades, I. García-Rodríguez, C. Benavides, and H. J. S. Alaiz-Moretón, "Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data," vol. 20, no. 4, p. 1214, 2020.

[41] N. K. Manaswi, "Rnn and lstm," in *Deep Learning with Applications Using Python*: Springer, 2018, pp. 115-126.

[42] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. J. I. t. o. n. n. Schmidhuber, and l. systems, "LSTM: A search space odyssey," vol. 28, no. 10, pp. 2222-2232, 2016.

[43] A. J. P. D. N. P. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," vol. 404, p. 132306, 2020.

[44] F. Landi, L. Baraldi, M. Cornia, and R. J. N. N. Cucchiara, "Working memory connections for LSTM," vol. 144, pp. 334-341, 2021.

[45] S. Yang, X. Yu, and Y. Zhou, "Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example," in *2020 International workshop on electronic communication and artificial intelligence (IWECAI)*, 2020, pp. 98-101: IEEE.

[46] D. Ageng, C.-Y. Huang, and R.-G. J. I. A. Cheng, "A Short-Term Household Load Forecasting Framework Using LSTM and Data Preparation," vol. 9, pp. 167911-167919, 2021.

[47] D. Arpit, B. Kanuparthi, G. Kerg, N. R. Ke, I. Mitliagkas, and Y. J. a. p. a. Bengio, "h-detach: Modifying the LSTM gradient towards better optimization," 2018.

[48] L. Yu, J. Qu, F. Gao, Y. J. S. Tian, and Vibration, "A novel hierarchical algorithm for bearing fault diagnosis based on stacked LSTM," vol. 2019, 2019.

[49] Z. Huang, W. Xu, and K. J. a. p. a. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015.

[50] T.-Y. Kim and S.-B. J. E. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," vol. 182, pp. 72-81, 2019.

[51] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with densley connected convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0-0.

[52] T. Mantoro and M. A. Ayu, "Multi-faces recognition process using Haar cascades and eigenface methods," in *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, 2018, pp. 1-5: IEEE.

[53] A. Priadana and M. Habibi, "Face detection using haar cascades to filter selfie face image on instagram," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019, pp. 6-9: IEEE.

[54] C. Rahmad, R. A. Asmara, D. Putra, I. Dharma, H. Darmono, and I. Muhiqqin, "Comparison of Viola-Jones Haar Cascade classifier and histogram of oriented gradients (HOG) for face detection," in *IOP conference series: materials science and engineering*, 2020, vol. 732, no. 1, p. 012038: IOP Publishing.

[55] A. B. Shetty and J. J. G. T. P. Rebeiro, "Facial recognition using Haar cascade and LBP classifiers," vol. 2, no. 2, pp. 330-335, 2021.

[56] D. T. P. Hapsari, C. G. Berliana, P. Winda, and M. A. Soeleman, "Face detection using Haar cascade in difference illumination," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 555-559: IEEE.

[57] C. Shorten and T. M. J. J. o. b. d. Khoshgoftaar, "A survey on image data augmentation for deep learning," vol. 6, no. 1, pp. 1-48, 2019.

[58] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 international interdisciplinary PhD workshop (IIPhDW)*, 2018, pp. 117-122: IEEE.

[59] P. Chlap *et al.*, "A review of medical image data augmentation techniques for deep learning applications," vol. 65, no. 5, pp. 545-563, 2021.

[60] Q. Zheng, M. Yang, X. Tian, N. Jiang, D. J. D. D. i. N. Wang, and Society, "A full stage data augmentation method in deep convolutional neural network for natural image classification," vol. 2020, 2020.

[61] J. Yoo, N. Ahn, and K.-A. Sohn, "Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8375-8384.

[62] D. Liang, H. Liang, Z. Yu, and Y. J. T. V. C. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," vol. 36, no. 3, pp. 499-508, 2020.

[63] G. J. J. a. p. a. Chowdary, "Class dependency based learning using Bi-LSTM coupled with the transfer learning of VGG16 for the diagnosis of Tuberculosis from chest x-rays," 2021.