**Deep Learning approach to Facial Emotion Detection using Transfer Learning**

1. **INTRODUCTION**

As fascinating as computer systems are, they lack emotional states and thus cannot communicate their state facially, which humans can. Emotion is a mental state connected to the nervous system(Gaddam *et al.*, 2022). It plays a significant role in human lives as it communicates peoples’ perception, pleasure, displeasure, and sense of judgement, which are fundamental to social impact. While humans can seldom hide their emotions, their knowledge of it remains highly resourceful for humanity’s betterment. For instance, through the knowledge of patients’ emotions, psychologists can understand their pressing challenges towards healing them.

With the understanding of the relevance and prospects of knowing humans’ emotion, there has thus been a steep increase in the interest in emotional computing over the years, and several emotionally sensitive products have been developed, finding application in education, phycological wellbeing and many more. According to Hertenstein *et al.* (2009), there are several ways through which the emotions of humans can be communicated. Human emotions can be communicated through touch, vocal display, facial expression and many more. Among this list, facial expression has received unprecedented attention following the advancement in machine learning.

Over ages, Human Facial expressions have played significant role in non-verbal communication. It involves the movement of one’s facial muscles. For instance, a displeased father can squeeze the muscles of this faces to an unbecoming attitude of his son instead to yelling which can be embarrassing. Research has also showed that even animals often communicate their emotions through specific muscle movements(Tai and Chung, 2007).

In the bid to maximize the ability to capture, interpret and integrate facial expressions to systems, numerous techniques have been deployed. Products factories are now interested in developing products with facial recognition to improve quality services, investing massively in technologies such as smart homes, smart phones and many more(Kumar, Sampathila and Tanmay, 2022). Facial emotion recognition technology is thus widely used today which forms the motivation for this work.

Following these advancements and interests, it is important to have an effective system that can acutely and accurately detect humans’ emotion. The recent advanced methods of deep learning has been employed in several image processing problems; however, the novel and effective methods of transfer learning have not been well explored (Saxena, Khanna and Gupta, 2020). Hence, for this research, transfer learning models that can quickly identify seven emotions (anger, disgust, neutral, fear, sad and surprise) will be developed using a publicly available dataset on Kaggle, a subset of a research by Goodfellow *et al.* (2015). The dataset contains 28,709 training sets of images and 3,589 testing sets of images, labelled across the seven (7) emotions.

Ultimately, this report will answer the following questions about emotion detection systems.

1. Considering the several bits needed to represent image data, can we have a deep learning model that can recognize human emotion within the twinkle of an eye?

2. How effective can the modern techniques of transfer learning algorithms of VGG16 and GoogleNet be in detecting human emotions?

Through this project, we hope to showcase how the advancement in deep neural networks can be well maximized for image processing, particularly for facial emotion detection for any applications.

1. **BACKGROUND OF STUDY**

The plethora potentials of emotional computing have attracted several attention across research articles(Li *et al.*, 2023). While several had developed methods of image processing, some had taken to video analysis in real time. Despite the enormity of existing research articles, there still exist a significant gap between the accuracy achieved and the time taken to recognize human emotions.

Following the availability of training data and computational power, machine learning algorithms have been adopted, particularly the technology of neural networks which has uncommon ability to identify non-linear patterns. Gueorguieva, Georgiev and Valova (2003) explored the Multilayer Perceptron and Radial Basis Function (RBF) and obtained an accuracy of 73%. Wang *et al.*, (2004), who also used the same dataset obtained an accuracy of 92.4% using Adaboost.

Owing to disparities that may occur due to change in pose while detecting emotions, Ruan *et al.*, (2020) used the frontal face and non-frontal points of the faces. However, Tai and Chung (2007) earlier highlighted the significance of rather extracting selected features such as contour, displacement of eyebrows, eyes and mouth of the faces to reduce recognition time and improve accuracy. This stance is thus valid as the end-product of models is their integration with systems. In a review by Saxena, Khanna and Gupta (2020), it became obvious that the technology of transfer learning which has been validated to be efficient and fast for image processing has not been well explored for explored for facial emotion detection algorithms.

**2.1. Objectives**

The objectives of this project are therefore to:

1. develop a facial emotion detection model using transfer learning models of VGG16 and GoogleNet.
2. evaluate and compare the performance of the developed models on the metrics of accuracy, recall and F1 score.

**3.0 METHODOLOGY**

This work thus uses advanced methods of transfer learning, which are VGG-16 and GoogLeNet, designing and modifying them as necessary to suits the objectives of the work for facial emotion detection. Each network was designed and trained differently on the labelled dataset. After careful hyperparameter tuning, highly effective models are developed and tested on the test set and random images from the internet. Figure 3.1 provides a detailed description of the procedures that will be taken, starting from pre-processing the data to evaluating the models.

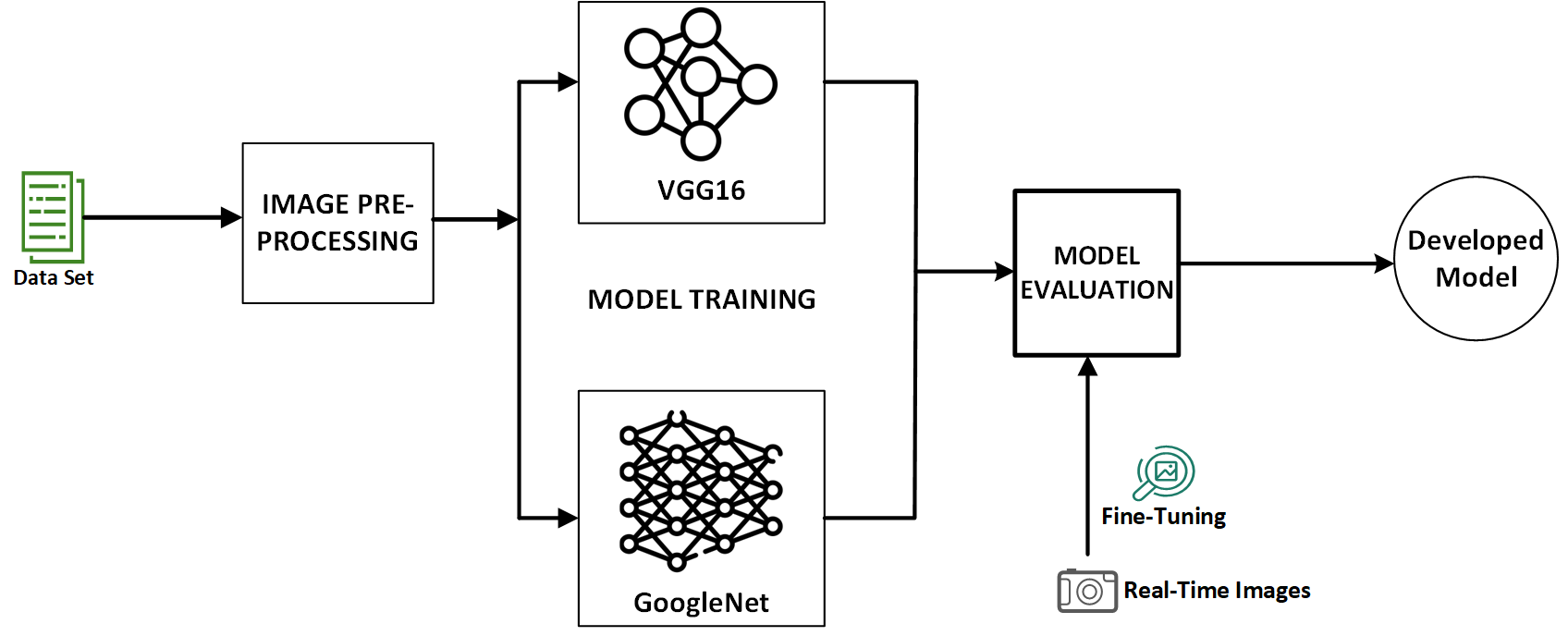


Figure 3.1 System Architecture

**3.1 Data Description and Overview**

The dataset consists of two folders labelled as train and test, containing 28,709 and 7178 images respectively. As shown in the four mages in Figure 3.2, each image is 48 by 48-pixel greyscale showing the emotion of the human. The two sets are labelled based on seven (7) kinds of emotion (angry:0, disgusted:1, fearful:2, happy:3, sad:5, surprised:6). Figure 3.2a and Figure 3.2b shows the distribution of the images across the 7 classes on the training and testing set respectively.

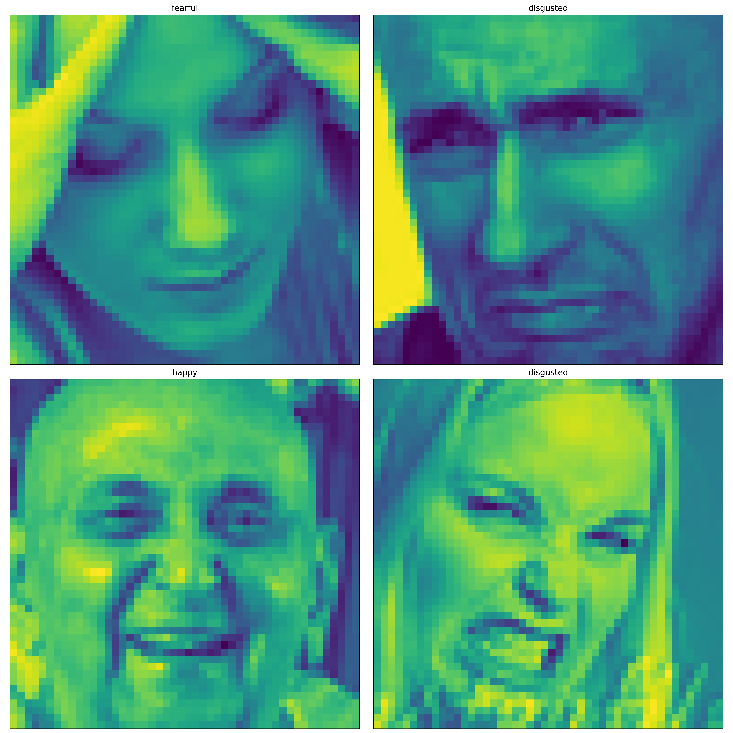
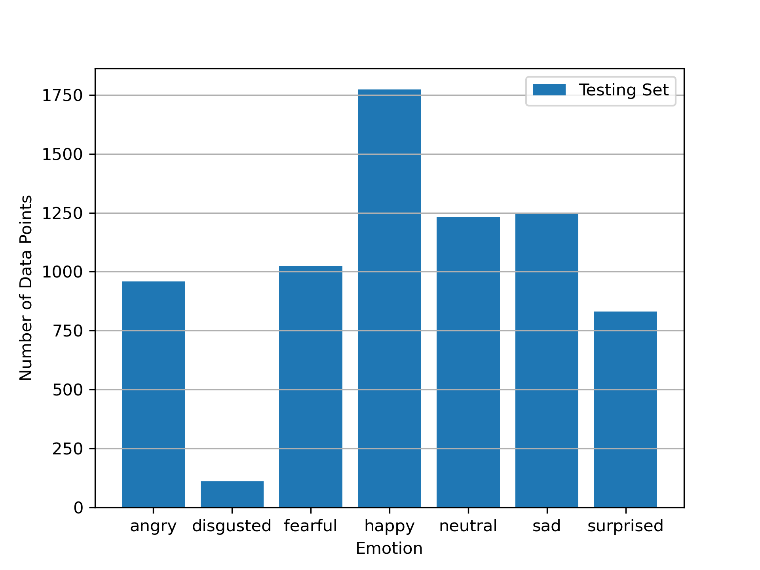
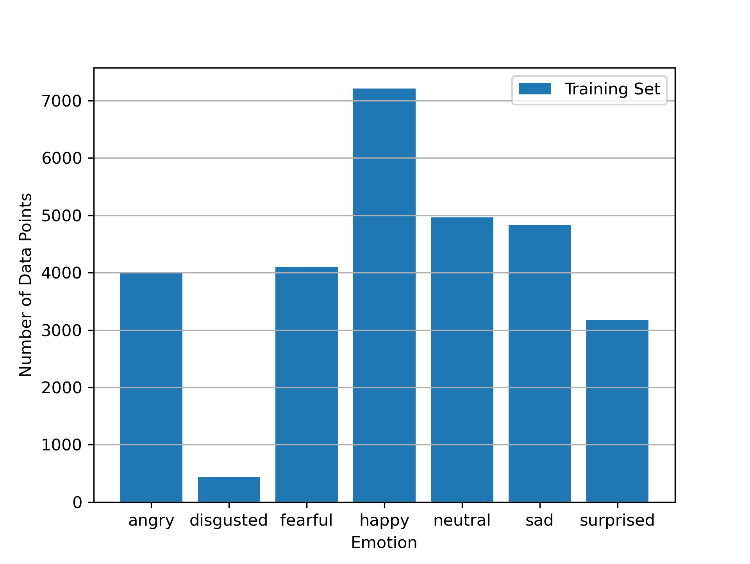


Figure 3.2 Image Overview



1. Training Set Data Distribution (b) Test Set Data Distribution

Figure 3.3 Image Distribution

**3.2 Image Processing**

It is necessary to clean the images before feeding them into the models as this will facilitate the system's processing of the features to be extracted from the data and also improve the efficiency. An image generator (a pipeline) was created for the testing and training set to assure more picture processing. After the data are extracted from each image (training and testing set), one of the crucial operations that is carried out is the normalization of the values. These image values were scaled from a range of 0-255 to 0-1.

**3.3 Modelling using VGG-16**

The VGG network design is a multilayered neural network in which features are automatically extracted at each layer and used to train the linking weights between each neuron. As the output generated is a one-hot vector with an output equal to 1 for whatever class of emotion the network creates, the classifier, which is the final layer, is configured using the softmax layer.

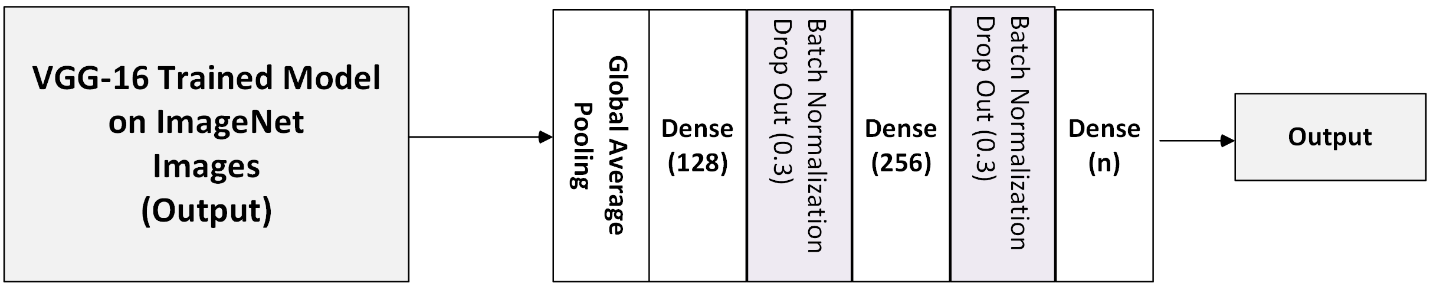


Figure 3.4 Modified VGG16 architecture adopted

Additional layers were added to the network to make sure the model accurately identified facial emotions in addition to identifying images in general which VGG16 has been trained on. Figure 4.1 displays each added layer's component. Batch Normalization and dropout were added to speed up training and avoid overfitting, which subsequently make learning easier because too many neurons put more stress on the network.

**3.4 Modelling using GoogleNet**

Using the tensorflow and keras libraries, the GoogLeNet architecture was set up and modelled in accordance with the layout depicted in Figure 3.5. This network is incredibly intricate. It’s developer claims that it was constructed with numerous deep layers to address the issue of overfitting. The model was trained using the training faces’ images.

A picture containing screenshot, line, diagram

Description automatically generated

Figure 3.5 GoogLeNet Architecture adopted

**4.0 EXPERIMENTAL SETUP**

The available dataset by was downloaded on a personal computer with window operating system. The 35,887 images were preprocessed and normalized between the 0 and 1 range for the algorithm to train. Training locally took more than expected period of time, hence, the google-collaboratory was employed that model may run without so much monitoring on the google server. In addition, the tensorflow which is a major library that helps us set up the VGG16 and the googleNET was installed both locally and on the google server. In the bid to obtain efficient results, the hyperparameters were tuned to excellence, with the performance of the model further evaluated. In addition, dropout and batch-normalization and regularizers were added to the network to avoid overfitting.

**5.0 RESULTS**

This section describes the results obtained upon using the processed data to train the two (2) models. The accuracies of the training data and validation data which were obtained by the dividing the data in the ratio 80:20 is provided is provided. Upon successful modelling, four images were obtained from the internet and tested on the two (2) developed models (GoogleNET and VGG16). The confusion matrix, which shows the performance of each model on every class of emotion is also provided. In addition, the total time taken to train the developed models were also noted and compared in this section.

**5.1. GOOGLNET**

Figure 3.4 and Figure 3.5 show the time taken to train the GoogLeNet model on the data and the accuracies obtained from training the training set and the validation set. Figure 3.4 shows it takes about 12,000 seconds to train the data over 30 epochs which gave the result obtained in Figure 3.5 where the training accuracy increased gradually to 63% while the validation accuracy increased to 55%.

**A graph with a line drawn on it

Description automatically generated**A graph with orange lines

Description automatically generated

Figure 3.4: Time Taken to Train GoogLeNet. Figure 3.5: Accuracy Result of GoogLeNet.

Upon evaluation of the model by estimating the performance on each class, the accuracy of each class is thus provided in the Figure 3.6. The confusion matrix shows the GoogLeNet model performs well on the following classes (‘Angry’, ‘happy’ and ‘sad’), while a low performance were experienced with other classes

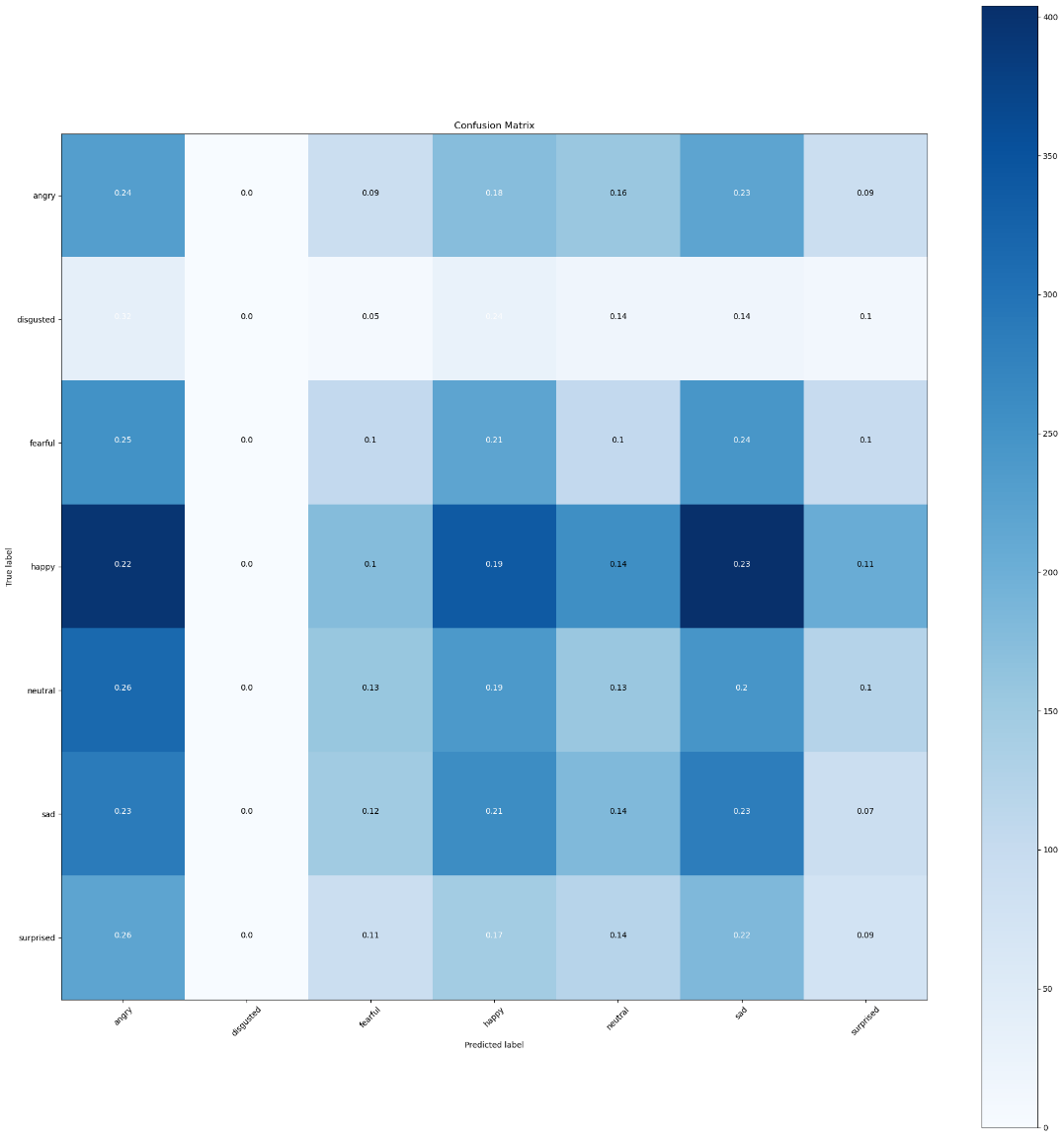


Figure 3.5: Confusion Matrix of GoogLeNet.

**5.2. VGG16**

Figure 3.6 and Figure 3.7 show the time taken to train the VGG16 model on the data and the accuracies obtained from training the training set and the validation set. Figure 3.8 shows it takes about 200,000 seconds to train the data over 30 epochs which gave the result obtained in Figure 3.5 where the training accuracy increased gradually to 98% while the validation accuracy increased to 63%.

A graph with a line

Description automatically generatedA graph with blue line and orange line

Description automatically generated

Figure 3.6: Time Taken to Train VGG16. Figure 3.7: Accuracy Result of VGG16.

Upon evaluation of the model by estimating the performance of each class, the accuracy of each class is thus provided in Figure 3.6. The confusion matrix shows the VGG16 model performs well on the following classes (‘Angry’, ‘happy’ ,’sad’, and ‘neutral’), while a low performance was experienced with other classes.

A blue squares with black text

Description automatically generated

Figure 3.8: Confusion Matrix of VGG16.

The results obtained showed that VGG16 model outperformed the GoogLeNet model by attaining a training accuracy of 98% and a validation accuracy of 63%, which indicate a subtle overfitting which can be corrected by addressing the unseemly biasness among the seven (7) classes of the images.

Table 1. Results using Internet Images

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/N** | **Original Image** | **Original Label** | **Prediction by GoogLeNet** | **Prediction by**  **VGG16** |
| 1 | A person with his mouth open  Description automatically generated | Surprised | Angry | Surprised |
| 2 | A close up of a person  Description automatically generated | Sad | Sad | Sad |
| 3 | A person with curly hair smiling  Description automatically generated | Happy | Neutral | Surprised |
| 4 | A close-up of a person  Description automatically generated | Angry | Sad | Angry |

Table 1.0 shows the results obtained upon testing randomly obtained images of people’s emotion from the internet. The images obtained are four (4) in number representing a surprised person, a sad person, a happy person and an angry person. The images were preprocessed and tested on each of the models to identify the emotion of each person. The GoogLeNet correctly predicted one out of the four images while the VGG16 correctly predicted three (3) out the four images correctly predicted, confirming its supremacy.

**6.0 CONCLUSION**

In this research, the novel method of deep learning (transfer learning) which maintains the balance between accuracy and processing time was adopted to recognize humans’ emotions from their images. The two models adopted (GoogLeNet and VGG16) showed a reckonable result, with the VGG16 outshining, attaining a training accuracy of 98% while the GoogLeNet lags with an accuracy of 63%, which was also validated by testing the models on randomly download images of humans from the internet. The result showed a significant result to reckon with, however, a subtle overfitting was recorded in both models which it is believed to have results from the biasness in the data distribution across the considered seven (7) emotions. Also, the VGG16 models took a longer time compared to GoogLeNet. It can therefore be said that transfer learning is highly resourceful to emotion detection which can be deployed across various systems.

**References**

Gaddam, D.K.R. *et al.* (2022) ‘Human Facial Emotion Detection Using Deep Learning’, *Lecture Notes in Electrical Engineering*, 783, pp. 1417–1427. Available at: https://doi.org/10.1007/978-981-16-3690-5\_136.

Goodfellow, I.J. *et al.* (2015) ‘Challenges in representation learning: A report on three machine learning contests’, *Neural Networks*, 64, pp. 59–63. Available at: https://doi.org/10.1016/j.neunet.2014.09.005.

Gueorguieva, N., Georgiev, G. and Valova, I. (2003) ‘Facial expression recognition using feedforward neural networks’, *Proceedings of the International Conference on Artificial Intelligence IC-AI 2003*, 1, pp. 285–291.

Hertenstein, M.J. *et al.* (2009) ‘The Communication of Emotion via Touch’, *Emotion*, 9(4), pp. 566–573. Available at: https://doi.org/10.1037/a0016108.

Kumar, S., Sampathila, N. and Tanmay, T. (2022) ‘Wavelet based machine learning models for classification of human emotions using EEG signal’, *Measurement: Sensors*, 24, p. 100554. Available at: https://doi.org/https://doi.org/10.1016/j.measen.2022.100554.

Li, Z. *et al.* (2023) ‘Facial Micro-Expression Recognition Using Double-Stream 3D Convolutional Neural Network with Domain Adaptation’, *Sensors*, 23(7). Available at: https://doi.org/10.3390/s23073577.

Ruan, S. *et al.* (2020) ‘Multi-Pose Face Recognition Based on Deep Learning in Unconstrained Scene’, *Applied Sciences*, 10(13). Available at: https://doi.org/10.3390/app10134669.

Saxena, A., Khanna, A. and Gupta, D. (2020) ‘Emotion Recognition and Detection Methods: A Comprehensive Survey’, *Journal of Artificial Intelligence and Systems*, 2(1), pp. 53–79. Available at: https://doi.org/10.33969/ais.2020.21005.

Tai, S.C. and Chung, K.C. (2007) ‘Automatic facial expression recognition system using Neural Networks’, in, pp. 1–4. Available at: https://doi.org/10.1109/TENCON.2007.4429124.

Wang, Y. *et al.* (2004) ‘Real time facial expression recognition with adaboost’, *Proceedings - International Conference on Pattern Recognition*, 3(January 2004), pp. 926–929. Available at: https://doi.org/10.1109/ICPR.2004.1334680.