International Journal on Engineering Technology (InJET)

www.kec.edu.np/kec-journal/
Volume 2, issue no. 1, Dec 2024, Page 227-236
Received Date: Aug 17, 2024
Accepted Date: Sept 23, 2024
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v2i1.72579

CNN-BiLSTM based Facial Emotion Recognition

Er. Alina Lamichhane[1,*], Er. Gopal Karn[2]

[1]*Software Engineer, Auxfin Development Nepal, Kupondole, Lalitpur, Nepal, alinalamichhane05@gmail.com, +977-9864421255*
[2]*Lecturer, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, gopal@kec.edu.np, +977-9846906061*

Abstract

Human emotions play an important role as they let articulate themselves without any words. Emotion Recognition involve a lot of information about facial expression, body language, tone, and pitch of voice etc. Among all this, Facial expression also plays an important role in interaction between humans and machines and significant amount of research had been done in face emotion recognition. Traditionally feature extraction from the image were done manually but over some past years different Machine Learning (ML) algorithms and Neural Network (NN) had been used for face emotion recognition. In this paper, Hybrid Neural Network CNN-Bi LSTM is used to extract the feature from facial image and detect emotion. To proceed this publicly available FER2013 dataset is used. The model is trained with greyscale image to classify seven emotions such as happy, sad, disgust, angry, fearful, surprise and neutral. The CNN component extracts spatial features, while the BiLSTM layer processes these features to capture temporal dependencies. The model achieves an accuracy of 79.4% when classifying all seven different emotions. However, when limited to detecting three emotions (happy, sad, neutral), the accuracy improves to 89.0%, demonstrating the model's potential for focused emotion recognition tasks.

Keywords: Hybrid Neural Network, CNN-BiLSTM, Image Processing, Face emotion Recognition, FER2013 Dataset

1. Background and Justification of the Concept

Facial emotion recognition (FER) has gained significant attention in recent years as it plays a crucial role in understanding human emotions, behaviors, and mental states. Accurate recognition of these expressions is essential for a wide range of applications, including human-computer interaction, psychological analysis, and surveillance. The ability to accurately detect and classify emotions from facial expressions has the potential to revolutionize many industries. The accurate recognition of facial emotions from images is a challenging task due to the complexity and variability of human facial expressions. Traditional methods often struggle to capture the nuances of different emotions, leading to suboptimal performance. Therefore, there is a need for advanced and robust models that can effectively recognize and classify facial emotions with high accuracy and efficiency. The research aligns with the objective of developing and evaluating deep learning models for facial emotion recognition.

Among the most effective techniques in FER are deep neural networks. Convolutional Neural Networks (CNNs) have shown great success in extracting spatial features from facial images, making them a popular choice for image-based tasks. However, CNNs are limited when it comes to capturing temporal dependencies and sequential information, which can be crucial for understanding the dynamic aspects of facial expressions. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are designed to process sequential data and address this limitation. LSTMs help maintain long-range dependencies by preserving information over time, mitigating issues like vanishing gradients that are common in standard RNNs. Bidirectional LSTM (BiLSTM) networks further enhance this capability by processing information in both forward and backward directions, providing a more comprehensive understanding of temporal relationships in data. This paper introduces a CNN-BiLSTM hybrid deep neural network for facial emotion recognition, designed to address gradient vanishing issues and improve memory retention. The CNN

*Corresponding Author

component efficiently extracts key features, while the BiLSTM layer identify emotions by processing these features. This approach aims to overcome the limitations of standalone CNNs in handling sequential data and to enhance memory retention for improved recognition accuracy. The system's performance is evaluated using different matrices, demonstrating its ability to identify seven universal emotions: disgust, anger, fear, happiness, sadness, neutral, and surprise. Unlike previous studies that utilize smaller datasets, such as JAFFE, this research emphasizes the importance of larger datasets to ensure high accuracy and practical efficiency in real-world applications. Facial expression recognition plays a pivotal role in areas such as intelligent human-machine interfaces, visual surveillance, teleconferencing, and real-time animation. Given the complexity of emotion detection, developing an algorithm that can autonomously and effectively perform this task is crucial.

The FER2013 dataset, which includes seven universal emotions anger, disgust, fear, happiness, sadness, surprise, and neutral is used to train and evaluate the proposed model. Although class imbalance within the dataset presents a challenge, our research focuses on developing a robust architecture capable of performing well across all emotions. The hybrid model's performance is compared with a standalone CNN model to evaluate its effectiveness. Thus, the key motivation behind this research is to explore real-time face emotion recognition through facial expressions, providing initial testing towards facilitating intelligent human-computer interaction

The relevance of this research lies in its potential to enhance human-computer interaction, mental health diagnosis, customer sentiment analysis, teleconferencing and real-time animation. Accurate facial emotion recognition can enable more intuitive and responsive human-computer interfaces, leading to improved user experiences. In the healthcare domain, it can aid in the early detection of mental health disorders by analyzing facial expressions. Moreover, in marketing and customer service, emotion recognition can provide valuable insights into customer sentiment and preferences. Thus, the key motivation behind this research is to explore real-time face emotion recognition through facial expressions, providing initial testing towards facilitating intelligent human-computer interaction

2.Related Works

The study of face detection and emotion recognition has been a highly active area of research with extensive literature available. Researchers are facing a significant challenge due to the lack of spontaneous expression data (Bettadapura, 2012). Capturing spontaneous expressions in images remains a major obstacle. Numerous efforts have been made to identify facial expressions, as (Mehendale, 2020). stated that facial expressions are a universal signal for conveying mood. In recent years, there has been considerable research on automatic facial emotion recognition (FER). Initially, traditional machine learning approaches were used for FER. The earliest approach for facial emotion recognition was based on distance urged, as discussed in (Khan, 2022). This approach utilized PCA (Principal Component Analysis) to detect facial action units and establish different emotions.

While classic facial recognition techniques using handcrafted features have shown significant success, researchers have increasingly turned to deep learning due to its powerful automatic recognition capabilities. Several new FER experiments have demonstrated the potential of deep learning techniques for improving detection (Mellouk, 2020). (Mehendale, 2020) proposed the use of deep CNN and supervised learning for facial emotion recognition by constructing a database and conducting corresponding ground-truth cross-validations. Deep CNN generally consists of three layers: (1) convolution layer, (2) pooling layer, and (3) fully connected layer. In this paper, the initial keyframe is extracted from the input video, and then the background is removed before passing the image through different CNN layers. Subsequently, deeper neural networks such as 3D CNN, RNN, and CNN-RNN were also implemented. (Khan, 2022) noted the implementation of 3D CNN for recognizing various emotions from videos. Additionally, many studies were conducted using CNN as a base network. (Donahue, 2014) proposed Long-term recurrent convolutional networks for visual recognition and description. (Bui, 2021) introduced a hybrid deep learning approach for facial emotion recognition, combining Convolutional neural networks for feature extraction and Long Short-Term Memory (LSTM) models for recognizing facial expressions.

In the field of natural language processing, bidirectional recurrent neural networks gained popularity and were introduced. We have implemented a model combining CNN-LSTM and Bidirectional RNNs to improve accuracy in predicting facial emotions. Throughout these research and model implementations, the models were evaluated (trained and tested) on different databases such as CK+, JAFFE, BU-3DFE, FERC, and FED-RO. According to (Bui, 2021), the hybrid deep neural network CNN-LSTM architecture was trained and tested using the small dataset JAFFE.

Regarding the applications of automatic facial emotion analysis and recognition models and tools, they have been utilized in various fields such as robotics, medicine, driving assistance systems, and lie detection (S. Dwijayanti, 2022) (Daniel Nixon, 2022). In robotics, real-time implementation of automatic facial emotion recognition can be used in humanoid robots (S. Dwijayanti, 2022). Additionally, automatic facial emotion recognition can be highly beneficial in depression counseling therapy (Daniel Nixon, 2022).

3.Methodology

In this section, we highlight the hybrid model which consists two learning techniques: Convolution neural network (CNN) and Bidirectional Long short-term memory (BiLSTM). The equivalent Bidirectional LSTM receives each visual characteristic determined by a CNN and outputs a fixed or variable length vector representation. The output is subsequently sent to a module for sequence learning. Finally, the SoftMax activation function is used to construct the anticipated distribution.

3.1. Convolution Neural Network (CNN)

CNN is a deep learning model that processes images by assigning weights and biases to different features and then classifying the image. Its architecture mimics the neural connections in the human brain. CNN mainly consists of three layers. The primary component is Convolution Layer that extracts features from an image using filters (kernels). It performs operations like convolution to generate feature maps, which highlight essential features of the image. Next is Pooling Layer, this layer reduces the spatial size of feature maps to decrease computational complexity while retaining important features. Common pooling methods include max-pooling (selecting the maximum value) and average-pooling (calculating the average value). Fully Connected Layer is Responsible for image classification, this layer flattens the feature maps and applies activation functions to categorize the image into different classes. Additional operations within CNN include: Dropout that prevents overfitting by randomly ignoring certain neurons during training. Batch Normalization that speeds up and stabilizes training by normalizing inputs between layers. Activation Functions like SoftMax, ReLU, and ELU are used to make the final classification decisions. SoftMax, for example, is commonly used for multi-class problems as it assigns probabilities to each class.

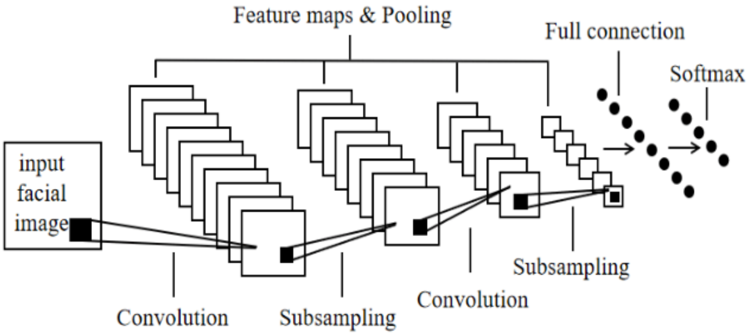

Figure 1. Procedure of Convolution Neural network *(Huang, 2019)*

3.2. Bidirectional Long Short-Term Memory (BI-LSTM)

Long short-term memory network is a special configuration of basic RNN network. A Recurrent neural network (RNN) is a class of artificial neural network which has the property of memorization as connections between the layers make a cycle. As it can remember the sequential data and produce output depending on the previous computation, it can be applied to many fields. However, when it comes to memorizing long term sequential data there is an issue in RNN. So, to gap “Long-Term Dependencies” modification to RNN using different gates results in Long Short-term Memorization (LSTM). These gates prevent gradient vanishing in conventional RNNs by being able to record both long-term memory and short-term memory along the time steps. "Forget gate," "Input gate," and "Output gate" are the names of the gates. The LSTM can also work in the inverse direction. By combining forward pass and backward pass of LSTM makes Bi-LSTM network. The bidirectional LSTM is considered as more accurate in storing data. Bidirectional LSTM in this paper is used to reconsider a part of previously trained features and it compares the image once in the forward and then in the backward direction which adds robustness to our model. The structure of Bidirectional LSTM cells is shown in figure 2.

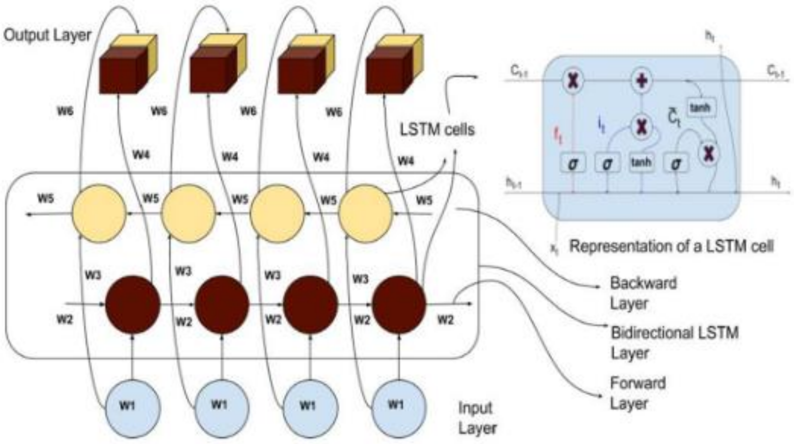

Figure 2. The Bidirectional LSTM cell *(Halder, 2020)*

The equation of the forget gate layer is given as,

$$f_t = \sigma\left(W_f[h_{t-1}, x_t], b_f\right) \qquad \text{Equation 1}$$

The next layer is called the input gate layer i_t, in this layer, the remember state data are retrained with the new features.

$$i_t = \sigma\left(W_f[h_{t-1}, x_t], b_i\right) \qquad \text{Equation 2}$$

The output from the forget gate layer is multiplied by the previous LSTM cell's cell state vector (c_{t-1}). This result is combined with the output from the input gate layer, which is multiplied by the hidden state vector from the previous step after passing through a "tanh" function. Together, they form the cell state vector for the next LSTM cell. This new cell state, after going through a "tanh" function, is then multiplied by the previous hidden state vector (h_{t-1}), which has passed through a "sigmoid" function, to generate the hidden state vector for the next LSTM cell (h_t).

In the final layer C'_t, a portion of the features from the previous state is combined with the newly modified features of the current cell, and the sum is passed to the next state.

$$C'_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \qquad \text{Equation 3}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C'_t, \qquad \text{Equation 4}$$

$$O_t = \sigma((W_o \cdot [h_{t-1}, x_t] + b_o)), \qquad \text{Equation 5}$$

$$h_t = O_t \cdot \tanh(C_t) \qquad \text{Equation 6}$$

where, x_t is an input vector to the LSTM unit and b_f, b_i and b_o are the weight vectors for the forget gate layer, input gate layer and the output gate layer, respectively.

3.3. Hybrid Deep Learning Approach (CNN-Bi LSTM)

The Hybrid deep learning approach was originally proposed by (Donahue, 2014). This model combines the advantages of both CNN and LSTM models. The Structure of Hybrid deep learning has four phases which is shown in figure 2. The first phase of the model has an input layer for an image of size 48 × 48 (height and width in pixels) and the second phase is CNN layer where first convolution is performed on the input and different operations are performed such as pooling, normalization and dropout. As the final output

of the CNN layer is produced, it is first flattened and fed to a bidirectional LSTM layer. The fixed or variable length vector representation as output of BiLSTM is inputted to an output or dense layer which can be defined by below equation (1) for classification of face images to various emotion classes as happy, sad, angry, disgust, neutral, fear and surprise. The entire architecture of our proposed model is illustrated by figure 3.

$$y = W_d X_t + B_d$$ Equation 7

where,
y = output of the output layer i.e. facial emotion of target image
X_t = the vector learned from CNN-Bi LSTM model
W_d, B_d = the weight and bias of CNN-Bi LSTM model

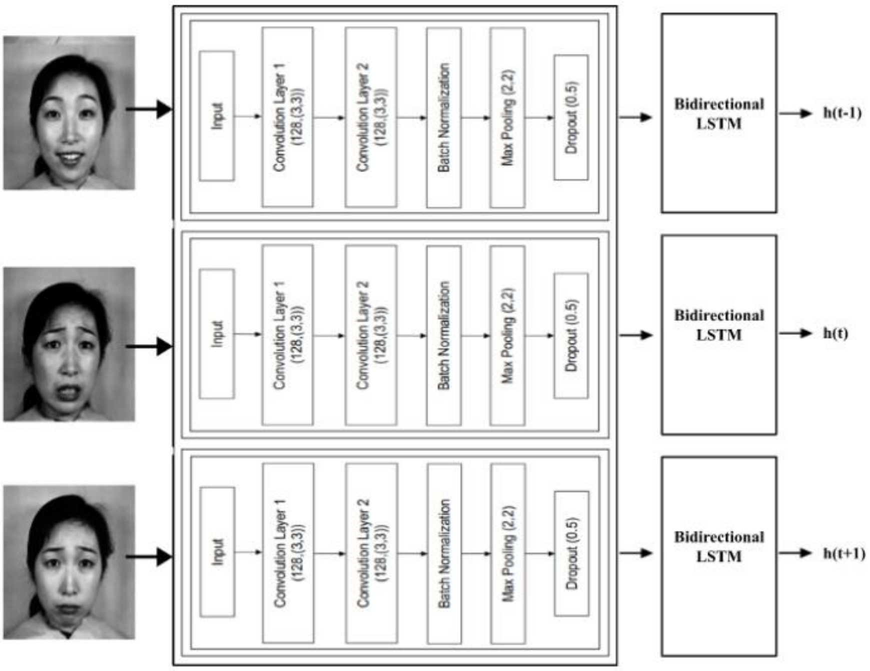

Figure 3. Hybrid Deep learning network

4.Experiments and Results

4.1. Dataset

Common dataset, FER2013 Dataset is used to train and evaluate the effectiveness of the suggested model. Due to several difficulties with this dataset, the data must be prepared for input to CNN. Images must be transformed into arrays since the model's input should be a collection of numerical values.

Table 1. Characteristics of FER2013 Dataset

Dataset	**Number of training**	**Number of test**	**Emotions**

	images	**images**	
FER2013	28,709	3,589	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral

Table 2. Descriptive characteristics of FER2013 Dataset

Expression	**Total**	**Anger**	**Disgust**	**Fear**	**Surprise**	**Neutral**	**Sad**	**Happy**
Tag	-	0	1	2	5	6	4	3
Train Quantity	28709	3995	436	4097	3171	4965	4830	7215
Test Quantity	3589	958	111	1024	831	1233	1247	1774

4.2. Result of CNN model

Initially, the CNN model was trained using a learning rate (LR) of 0.1, batch size of 32, training-validation ratio of 0.5, and the Adam optimizer. However, the results from this configuration were unsatisfactory, likely due to a high learning rate causing instability during optimization, leading to poor convergence.

To address this, the parameters were fine-tuned as follows: the learning rate was reduced to 0.001, dropout was introduced with a rate of 0.1 to mitigate overfitting, the batch size was decreased to 16 to allow for more frequent updates to the weights, and the training-validation split was adjusted to 0.8 to allow more data for training. This configuration aimed at achieving better generalization and convergence stability.

The model was then trained for 100 epochs with these modified parameters. The results of the model with modified parameters is illustrated in Figure 4.

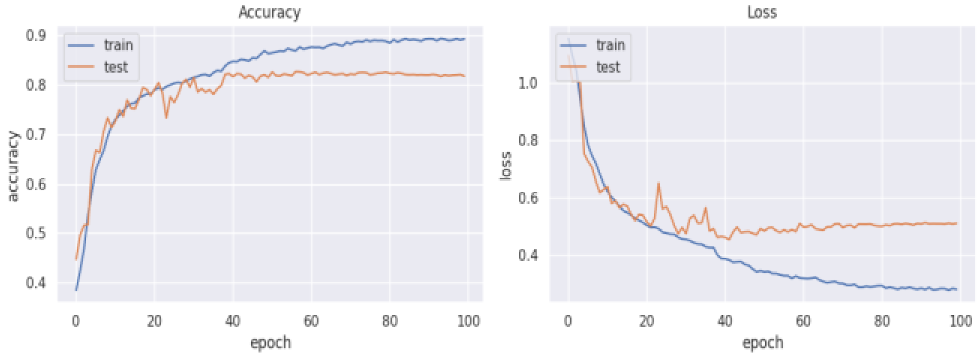

Figure 4. Accuracy and loss plot for results of CNN model after training for 100 epochs

4.3. Result of CNN-BiLSTM model

To assess the performance of our proposed hybrid model, the CNN-BiLSTM, we conducted two sets of experiments. In the first set of experiment, we tested the CNN-BiLSTM model across all seven emotions from the FER2013 dataset (happy, sad, angry, fear, surprise, disgust, and neutral). The accuracy and loss plots for this experiment are illustrated in Figure 5, showing how the model performed over multiple epochs.

For the second set of experiment, we focused specifically on a subset of three emotions: happy, sad, and neutral. This was done to examine whether reducing the classification complexity could lead to a more significant improvement in model performance. The accuracy and loss plot for the three-emotion set is displayed in Figure 6.

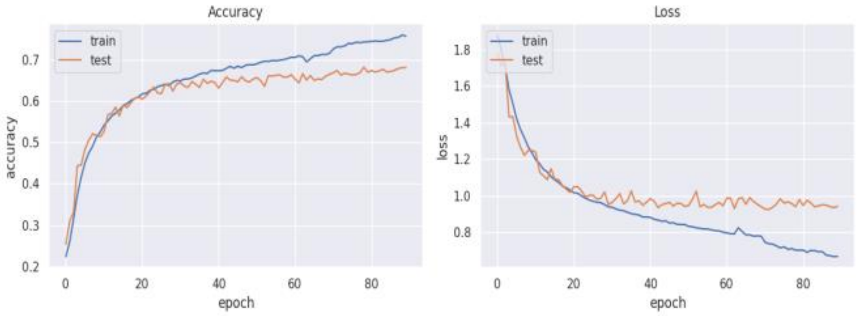

Figure 5. Accuracy and loss plot for results of CNN-Bi LSTM model for all seven emotions

s

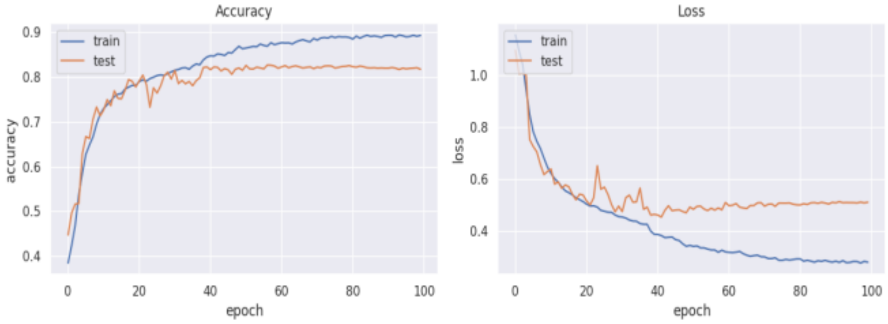

Figure6.Accuracy and loss plot for results of CNN-Bi LSTM model only for 3 emotion set (Happy, sad, neutral)

4.4. Comparison and evaluation of result

Upon comparison, the CNN-BiLSTM model takes approximately 278 seconds for a single step while the standalone CNN model averages 300 seconds per trail. Training with early stopping required between 80 and 100 epochs for CNN-BiLSTM model. In terms of accuracy, both models performed similarly on all seven emotions, with CNN-BiLSTM achieving an average accuracy of 79.6% and CNN achieving 79.3%. The average loss for the CNN-BiLSTM model was 0.62, while the CNN model had a slightly lower loss of 0.61. This marginal improvement is likely influenced by the imbalanced nature of the dataset, where minority emotions like "Disgust" contribute less to overall accuracy due to fewer samples, thereby limiting the potential gains from the BiLSTM's sequential learning capabilities.

However, when tested on three specific emotions (happy, sad and neutral), the CNN-BiLSTM model showed improved performance, achieving an average of 89.0% with a corresponding loss of 0.53. This improvement is attributed to the BiLSTM's ability to capture more subtle variations in the temporal and spatial features of emotions.

Despite the modest overall improvement, the addition of the BiLSTM layer offers certain advantages. BiLSTM layers are particularly effective at capturing sequential dependencies, which is crucial for tasks like facial emotion recognition where temporal dynamics and subtle facial changes are important. While the CNN model may suffice for static image classification, the BiLSTM architecture is designed to improve generalization, especially in real-world applications such as video-based emotion recognition or dynamic facial expressions.

The increase in computation time is a recognized trade-off, but the added complexity provides a stronger foundation for handling sequential data and potentially enhances the model’s robustness to varied facial expressions in a real-time or video-based context. Future work will focus on optimizing the hybrid model to reduce computational costs while maintaining its ability to capture complex temporal patterns.

4.5. Testing of CNN-BiLSTM model

Some random images of sad and neutral face emotion are taken using NumPy random choice function. These facial emotions are then tested and plotted with true emotion set and predicted emotion set. The test result is illustrated in figure 6 below.

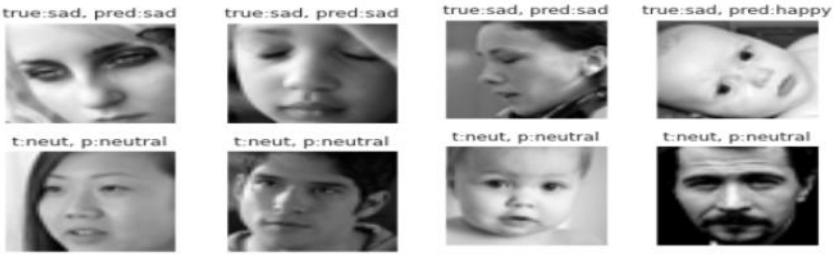

Figure 7. Predicted Facial Emotion with True Facial Emotion.

4.6. Real-time Testing of CNN-BiLSTM model

As a practical demonstration of the CNN-BiLSTM model, real-time face emotion recognition was tested using a webcam. The Haar cascade classifier was used to detect and isolate facial regions from the live video feed, while the trained CNN-BiLSTM model was responsible for identifying emotions from the detected faces. Each detected face was highlighted, and the corresponding emotion was displayed in real-time as shown in figure 8.

While this real-time testing provided a qualitative demonstration of the model’s potential for real-world applications, detailed performance metrics such as accuracy, precision and processing speed were not calculated at this stage. Preliminary observations showed that the model was able to detect and classify facial emotions in real-time under controlled lighting conditions. However, its robustness to varying conditions such as different facial angles, expressions, and lighting levels could not be fully evaluated.

Future work will focus on a more rigorous real-time evaluation, incorporating quantitative metrics like accuracy, precision, recall, and speed to better assess the model's performance in dynamic, uncontrolled environments. This will help validate the model’s suitability for practical applications in human-computer interaction systems.

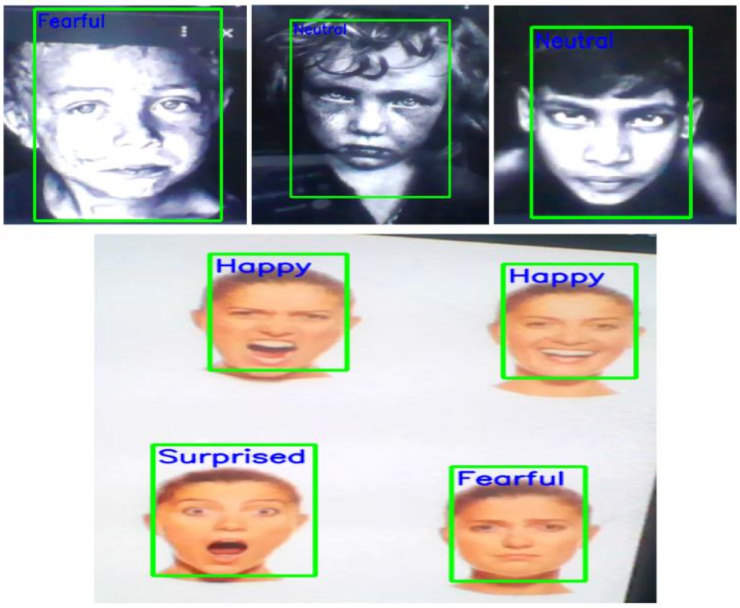

Figure 8. Real time prediction of facial emotion

5.Conclusions

In conclusion, this study developed and evaluated two models, CNN-BiLSTM and CNN, to extract facial features and recognize emotions using the FER2013 dataset. The dataset comprises seven emotions: happy, sad, angry, fear, surprise, disgust, and neutral. However, as shown in Table 1 the suffers from a significant imbalanced distribution with certain emotions, such as "Disgust" being underrepresented, with only 436 training samples compared to more frequent emotions like "Happy" (7215 samples). This class imbalance poses a challenge for the models, particularly when learning from minority classes, which may affect the ability to fully capture the features associated with these emotions.

To maintain the comparability of results between the standalone CNN and CNN-BiLSTM models, techniques such as data augmentation or weighted loss functions were not applied in this study. However, this lack of mitigation likely contributed to the modest improvement in accuracy observed when using the CNN-BiLSTM model across all seven emotions. The class imbalance inherently limits the model's ability to learn from less-represented emotions like "Disgust." The small performance gap between the two models may also reflect that CNN is already quite effective at learning from static images, while the advantage of the BiLSTM layer designed to capture temporal dependencies was less pronounced due to the static nature of the data.

The CNN-BiLSTM model was designed with three phases: the first phase includes the convolution layers for feature extraction, the second phase involves the Bidirectional LSTM layer to capture any potential temporal relationships between facial features and the third phase include fully connected and output layers. A separate CNN model similar to the CNN-BiLSTM model but without the LSTM layers was also developed for comparison. When evaluated across all seven emotions, both models achieved a similar average accuracy of 79.4%, with an average loss of 0.61. For the three-emotion classification (happy, sad, neutral), the CNN-BiLSTM model achieved an improved performance, achieving an average accuracy of 89.0%, with a corresponding average loss of 0.53, likely due to the reduced complexity of the classification task.

While the CNN-BiLSTM model offers a slight performance gain in some scenarios, the added complexity may not fully justify its use for static image-based emotion recognition without addressing class imbalance. Future work could focus on improving model performance by incorporating techniques to mitigate dataset imbalance, such as data augmentation, weighted loss functions, or oversampling, to ensure better learning from underrepresented emotions. Additionally, exploring the hybrid model's potential in more dynamic, sequential contexts such as video-based emotion recognition could better demonstrate its advantages over standalone CNNs.

References

Bettadapura, V., 2012. Face Expression Recognition and Analysis: The State of the Art. *arXiv.*

Bui, H. &. T. L. M., 2021. *Facial Expression Recognition with CNN-LSTM.* s.l.:Research in Intelligent and Computing in Engineering.

Daniel Nixon, V. V. M. V. P. S. H. S. K. K., 2022. A novel AI therapy for depression counseling using face emotion techniques,. *Global Transitions Proceedings,* 3(1), pp. 190-194.

Donahue, J. &. H. L. &. G. S. &. R. M. &. V. S. &. S. K. &. D. T., 2014. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Halder, R. C. R., 2020. CNN-BiLSTM Model for Violence Detection in Smart Surveillance. *SN Computer Science,* 1(4).

Huang, Y. &. C. F. &. L. S. &. W. X., 2019. Facial Expression Recognition: A Survey. *Symmetry.*

Jeff Donahue, L. A. H. M. R. S. V. S. G. K. S. T. D., 2014. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2625-2634.

Khan, A., 2022. Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges. *Information.*

Mehendale, N., 2020. Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences.*

Mellouk, W. &. W. H., 2020. *Facial emotion recognition using deep learning: review and insights.* s.l., Procedia Computer Science.

S. Dwijayanti, M. I. a. B. Y. S., 2022. Real-Time Implementation of Face Recognition and Emotion Recognition in a Humanoid Robot Using a Convolutional Neural Network. *IEEE Access.*