International Journal of Communication, Science and Technology
vol. 1, no. 2 (March-April, 2025), pp.75-94

A COMPARATIVE STUDY ON FACIAL EMOTION RECOGNITION (FER) USING CNN, CNN+LSTM & DENSENET

Subhabrata Bhattacharjee

Research Scholar, Supreme Knowledge Foundation Group of Institutions (SKFGI)

Mainak Mitra

Research Scholar, Supreme Knowledge Foundation Group of Institutions (SKFGI)

Prof. Koyel Chakraborty

Assistant Professor, CSE Department, Supreme Knowledge Foundation Group of Institutions (SKFGI)

Dr. Amit Chakladar

Professor, Techno India University

ABSTRACT

The research evaluates three DL models for FER through their application to CNN, CNN-LSTM and DenseNet architectures. The models were tested on different facial expression datasets to evaluate their ability to detect human emotional cues. The CNN model functions as our initial spatial feature extraction tool while LSTM addition enables the network to detect temporal patterns in facial movements. The DenseNet network stands out by using feature reuse to maintain consistent gradient flow across its entire structure. Our experimental results demonstrate both the advantages and constraints of each model which leads to detailed understanding of their performance capabilities. The research outcomes enhance our comprehension of these architectures in emotion detection tasks while establishing essential foundations for future investigations into network optimization and real-time data processing.

Keywords: *FER, CNN, LTSM, CNN+LSTM, DenseNet*

1. INTRODUCTION

AI uses FER as a vital field to enable computers to detect human emotions through facial expression analysis. The technology operates across multiple fields which include mental health monitoring as well as HCI and security and customer service. The accurate detection of emotions through technology enables better human-machine communication which results in more user-friendly and responsive systems.

Facial emotion detection has received significant improvement through the application of DL methods. Hand-crafted features in traditional methods proved insufficient for precise expression recognition because they did not handle complex expressions effectively. DL models include CNNs and advanced architectures such as LSTM networks and DenseNet. They enable automatic feature extraction from images which results in improved recognition accuracy [10].

2. MOTIVATION

The research objective aims to examine the performance of CNN, CNN+LSTM (Proposed Model), and DenseNet-121 DL models to determine their respective advantages and disadvantages in facial emotion detection. The three models demonstrate different strengths in image analysis through CNNs and sequential dependency analysis through LSTMs and deep feature extraction through DenseNet. The performance of these models in emotion recognition depends on dataset size and training epochs as well as model complexity.

The main objectives are:

I. To compare the accuracy, recall, precision, F1-score, and AUC of CNN, CNN+LSTM, and DenseNet-121 in FER.
II. To analyze their strengths and limitations based on experimental results.
III. To explore challenges such as overfitting, underfitting, and computational cost.
IV. To suggest possible improvements and future directions for better emotion recognition models.

The evaluation of these models will provide understanding of the most suitable DL method for facial emotion recognition while revealing opportunities for additional model development.

3. RELATED WORKS

3.1 Previous Research on FER

- The study of FER has been extensively researched throughout numerous years. Researchers employed handcrafted features in earlier studies by choosing particular facial points for emotion detection. The early facial recognition systems employed Eigenfaces and Fisher faces methods yet these approaches failed to handle real-world variations which included lighting conditions and facial expressions and pose changes [1].
- Researchers began applying ML techniques to facial image pattern detection through feature extraction methods including HOG and LBP. The methods achieved better results than manual feature selection but their performance remained limited when dealing with complex emotions and real-time applications.
- DL introduced a transformation to FER which revolutionized its capabilities. DL models automatically discover patterns from extensive data sets without requiring manual feature extraction. The transition led to significant improvements in emotion recognition accuracy and reliability particularly when using extensive datasets including FER2013, CK+ and AffectNet. Researchers have investigated various neural network architectures including CNNs, RNNs and Proposed Models to boost performance [2].

3.2 DL Models Used in Past Studies

The ability of DL models to detect facial emotions has advanced substantially because they detect intricate patterns in facial expressions. The following list contains the most frequently used models which researchers employed in their studies:

3.2.1 CNNs

The use of CNNs extends to various image-based applications which include FER tasks. The networks perform automatic feature extraction from facial images by detecting their edges together with textures and shapes. The research demonstrates that CNNs deliver superior performance compared to conventional ML approaches in terms of accuracy and generalization capabilities. Researchers have tested various CNN architectures including VGG16, ResNet and MobileNet to enhance emotion recognition performance.

3.2.2 RNNs and LSTMs

Researchers have investigated the application of RNNs and LSTM networks because emotions tend to manifest across time sequences. These models track facial expression changes by processing image sequences rather than single frames. Research studies have utilized

CNN+LSTM models to detect micro-expressions. They are brief, involuntary facial expressions that reveal genuine emotions.

3.2.3 RestNet and Other Advanced Architectures

Research has investigated RestNet which establishes connections between all layers throughout the network. The model becomes more efficient and less prone to vanishing gradients because feature reuse and gradient flow are enhanced. Recent studies have demonstrated that Transformers and Vision Transformers (ViTs) show promising performance in emotion recognition tasks.

3.2.4 Proposed Models

Researchers have achieved improved results by integrating multiple DL models together. The combination of CNN models with LSTM layers enables the model to detect spatial and temporal features. Researchers have employed GANs (Generative Adversarial Networks) to create additional training data which enhances recognition accuracy [3-4].

3.3 Existing Challenges in FER

DL has made significant progress but FER continues to encounter multiple obstacles:

3.3.1 Variability in Facial Expressions

The way people show emotions depends on their cultural background as well as their personal traits and facial characteristics. The same facial expression of a smile can mean happiness for one person but nervousness for another person. The difficulty of model generalization occurs because of this variability between different individuals.

3.3.2 Occlusions and Lighting Conditions

The visibility of facial expressions becomes limited when people wear glasses or masks or place their hands over their mouths. The detection of facial features becomes challenging when models encounter poor lighting conditions. The accuracy of emotion recognition models suffers due to these factors.

3.3.3 Data Imbalance in Datasets

Publicly accessible datasets contain an unbalanced distribution of samples across different emotional expressions. The majority of available datasets contain numerous neutral and happy images but limited samples of fear and disgust expressions. The uneven distribution of data

samples in the training set results in biased models that excel at identifying some emotions but fail to recognize others.

3.3.4 Real-Time Performance

DL models face significant challenges when it comes to real-time execution particularly on smartphones and embedded systems. The high computational requirements of performance models create deployment challenges for real-world applications including emotion-aware chatbots and smart surveillance systems.

3.3.5 Privacy and Ethical Concerns

The practice of emotion recognition creates privacy issues because it involves the unauthorized collection and analysis of facial expressions from people. The accuracy of these models can be affected by biases present in training data which leads to ethical concerns [5].

4. METHODOLOGY & TECHNIQUES

4.1 Datasets Used in FER

Choosing proper datasets plays a vital role in developing and assessing DL models for facial emotion recognition. A suitable dataset should contain multiple facial expressions together with various lighting conditions and multiple subjects to enhance model accuracy and generalization. The research utilized three established datasets FER2013 and RAF-DB and the FER Dataset together with a dataset created specifically for this study. The use of multiple datasets enhanced facial expression diversity which resulted in a more robust and effective model for real-world applications.

4.1.1 FER2013 (Facial Expression Recognition 2013)

FER 2013 represents a commonly utilized dataset for FER tasks. The dataset was first presented in the Kaggle FER Challenge and consists of 35,887 grayscale images that measure 48×48 pixels and are labelled with seven different emotions:

- Angry
- Disgust
- Fear
- Happy
- Sad
- Surprise

- Neutral

The dataset is useful because it has a large number of images [11].

4.1.2 RAF-DB

The RAF-DB dataset consists of 29,672 real-world images obtained from the internet. The images in this dataset offer more realistic facial expressions because they were obtained from various angles and lighting conditions and backgrounds. The dataset is divided into:

- **Basic Emotion Set**: Includes the same seven emotions as FER2013.
- **Compound Emotion Set**: Includes mixed emotions like “happily surprised” and “sadly angry.”

The RAF-DB dataset provides superior image quality and more realistic emotional representations than FER2013 which makes it suitable for DL models. [12]

4.1.3 FER Dataset

This dataset consists of high-resolution images and is widely used in research related to emotion recognition. It contains both posed and spontaneous facial expressions, helping models learn to detect emotions in different situations.

4.1.4 Custom Dataset

To improve model performance, we created our own custom dataset by collecting images from multiple sources, including:

- Real-world images from social media and public datasets.
- Webcam-captured images under different lighting conditions.

This custom dataset helped address some of the limitations found in existing datasets by adding more variety and real-life conditions.

Image 1: Samples of custom dataset [15]

4.1.5 Combining Multiple Datasets

Instead of using each dataset separately, we combined FER2013, RAF-DB, and the FER Dataset into a single, larger dataset. This provided several advantages:

- **Increased Data Diversity** – More variations in facial expressions, lighting, and angles improved the ability of the model to generalize.
- **Balanced Class Distribution** – By mixing datasets, we ensured better representation of underrepresented emotions like "disgust" and "fear."
- **Better Model Performance** – The combined dataset helped the model recognize subtle emotional differences more effectively [6].

4.1.6 Challenges with Datasets

While using multiple datasets improved model performance, we also encountered some common challenges:

- **Imbalanced Classes** – Some emotions, like "happy" and "neutral," had more images than "disgust" or "fear," leading to biased predictions.
- **Labelling Errors** – Some images in FER2013 had incorrect labels, which could mislead the model.
- **Different Image Sizes** – Since each dataset had different image resolutions, pre-processing was required to ensure uniformity [7].

4.1.7 Limitations of Datasets

This dataset is useful because it has a large number of images, but it also has some limitations. Some images are blurry, poorly cropped, or misclassified, which can reduce model accuracy [11].

4.2 Data Pre-processing Steps for FER

Pre-processing is an important step in FER because raw images contain noise, varying lighting conditions, and different orientations. To improve model performance, several pre-processing techniques were applied before training. These steps include feature extraction, normalization, label encoding, data balancing using SMOTE, and data augmentation.

4.2.1 Feature Extraction

The first step in pre-processing is extracting features from facial images. Since DL models require a standard input format, all images are converted into grayscale and resized to 48×48 pixels. Grayscale images reduce computational complexity while preserving important facial features.

In our implementation, the function extract features () loads each image, converts it into an array, and reshapes it into (48, 48, 1) format. This ensures that the dataset is compatible with CNNs.

4.2.2 Normalization

After feature extraction, pixel values are normalized to a range between 0 and 1 by dividing each pixel by 255.0. This step ensures that the model learns efficiently without large variations in pixel intensity. Normalization also helps speed up training and prevents large gradients from affecting model convergence.

4.2.3 Label Encoding and One-Hot Encoding

Since emotion labels are in text format (e.g., "happy," "sad," "angry"), they need to be converted into numerical values for model training. Label encoding is applied first, assigning a unique number to each emotion class. After that, one-hot encoding is used to convert these numerical labels into binary vectors. This ensures that the model does not assume any numerical relationship between different emotions.

For example, we have seven emotion categories:

- Angry → [1, 0, 0, 0, 0, 0, 0]
- Disgust → [0, 1, 0, 0, 0, 0, 0]
- Fear → [0, 0, 1, 0, 0, 0, 0]
- Happy → [0, 0, 0, 1, 0, 0, 0]

- Sad → [0, 0, 0, 0, 1, 0, 0]
- Neutral → [0, 0, 0, 0, 0, 1, 0]

4.2.4 Handling Class Imbalance using SMOTE

One major challenge in FER is class imbalance. Some emotions (like "happy" or "neutral") are more commonly available in datasets, while others (like "disgust" or "fear") have very few samples. This imbalance can result in biased predictions because the model tends to favour the majority classes.

To solve this problem, SMOTE is used. SMOTE generates synthetic samples for minority classes, balancing the dataset and improving model performance. Before applying SMOTE, training images are flattened into 1D vectors so that SMOTE can process them correctly. After oversampling, the data is reshaped back to the original (48, 48, 1) format.

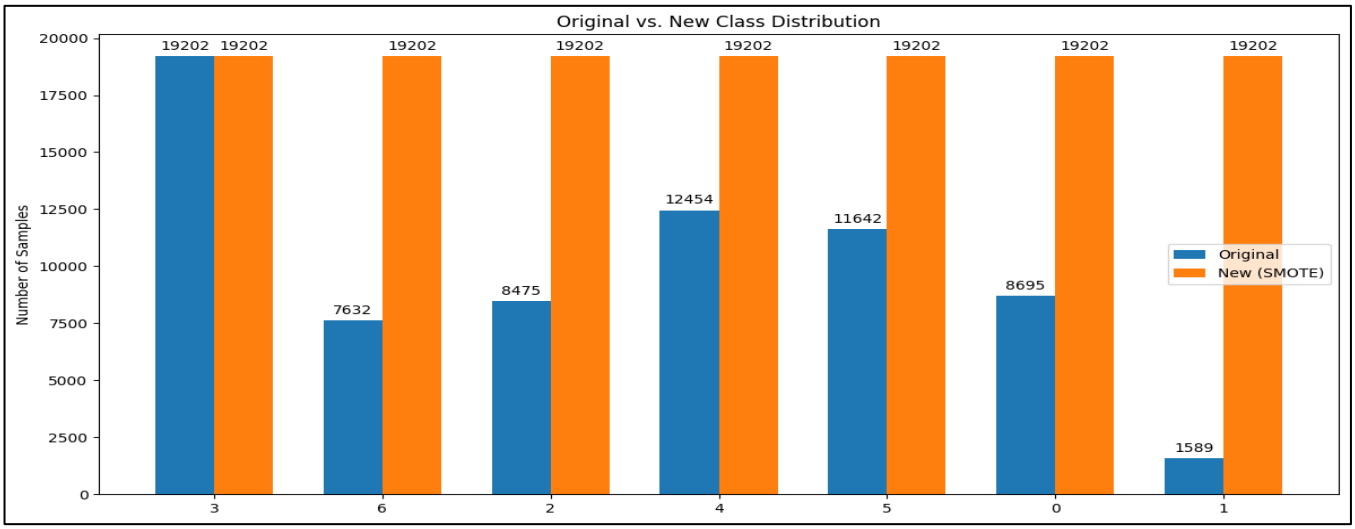

Figure 1: Original vs New Class Distribution

This effectively illustrate how SMOTE helps in balancing the dataset by increasing the representation of minority classes. [8]

4.3 Model Architectures

4.3.1 CNN

In this study, the first model used for FER is a CNN. CNN is widely used for image-related tasks as it can extract important features from images while maintaining spatial relationships.

4.3.1.1 Layers and Structure

The CNN model consists of multiple layers designed to learn patterns in facial expressions:

- **Convolutional Layers:** The model has three convolutional layers with 64, 128, and 256 filters, respectively. The network extracts facial features through these layers which detect edges and shapes and textures.
- **Activation Function:** The ReLU activation function operates in each convolutional layer to introduce non-linearity which enables the network to learn complex patterns.
- **Batch Normalization:** The application of batch normalization following convolutional layers serves to stabilize training processes and enhance convergence rates.
- **MaxPooling Layers:** MaxPooling functions as a reduction technique following convolutional layers to decrease feature map dimensions while maintaining essential features and minimizing computational requirements.
- **Dropout Layers:** The model uses dropout as an overfitting prevention technique which turns off specific neurons by chance during training at two different stages (0.3 and 0.4).
- **Fully Connected Layers:** The extracted features undergo flattening before passing through two dense layers containing 256 and 128 neurons.
- **Output Layer:** The output layer contains seven neurons that represent the 7 emotion classes with a softmax activation function for multi-class classification [14].

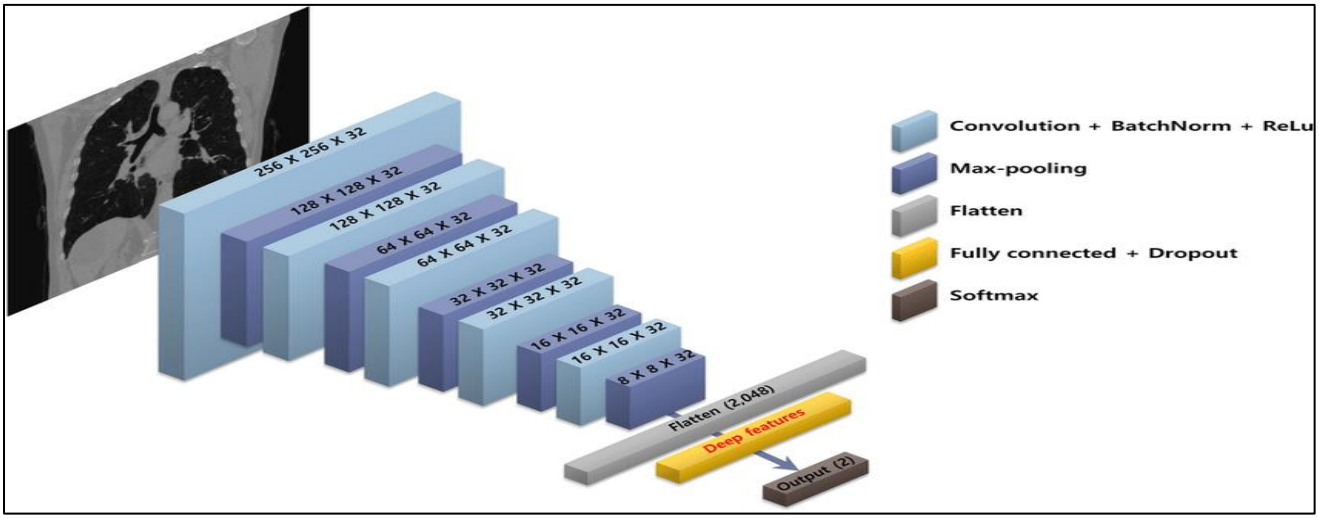

Figure 2: CNN Model [13]

4.3.1.2 Hyperparameters

The CNN model is trained using the following hyperparameters:

- **Optimizer:** Adam optimizer is used with a learning rate of 0.001, ensuring efficient weight updates.

- **Loss Function:** Categorical Crossentropy is used, as this is a multi-class classification problem.
- **Batch Size:** The model is trained with a batch size of 64 to balance memory usage and training speed.
- **Number of Epochs:** The model is trained for 100 epochs, allowing it to learn deep features over time.
- **Learning Rate Scheduler:** ReduceLROnPlateau is used to reduce the learning rate if the validation loss does not improve.
- **Early Stopping:** The training stops automatically if the validation accuracy does not improve for 10 consecutive epochs, preventing overfitting [13].

4.3.2 CNN+LSTM (Proposed Model)

LSTM was added to improve the model's ability to understand patterns over time. While CNN captures spatial features from facial images, it does not consider the relationship between these features. LSTM helps by processing these extracted features in a sequential manner, allowing the model to recognize subtle changes in facial expressions. This is useful for analyzing emotions that may appear differently depending on small variations in facial muscles. By combining CNN and LSTM, the model not only detects important facial features but also learns how these features are related over time, leading to better accuracy in emotion recognition.

4.3.2.1 Layers and Structure

4.3.2.1.1 CNN for Feature Extraction

- **Conv2D Layers (64, 128, 256 filters, kernel size 3x3):** Detects facial patterns like edges, textures, and shapes.
- **Batch Normalization:** Stabilizes training by normalizing activations.
- **MaxPooling2D:** Reduces spatial dimensions to focus on key features.
- **Dropout (0.3 - 0.4):** Prevents overfitting by randomly deactivating some neurons [8].

4.3.2.1.2 Transition to LSTM

- **Reshape Layer:** Converts CNN feature maps into a sequence format suitable for LSTM.
- **LSTM (128 units, activation='relu'):** Processes sequential features and learns temporal dependencies between different facial expressions.

4.3.2.1.3 Fully Connected Layers & Output

- **Dense (256, 128 neurons, L2 Regularization in 256-layer):** Extracts high-level emotion-related features.
- **Dropout (0.4, 0.3):** Prevents overfitting by adding randomness.
- **Softmax Output (7 classes):** Predicts the probability of each emotion [8].

4.3.2.1.4 Training Configuration

- **Optimizer:** Adam (learning rate 0.001) for efficient weight updates.
- **Loss Function:** Categorical Crossentropy for multi-class classification.
- **Callbacks:**
  - **ReduceLROnPlateau:** Lowers learning rate if validation loss stops improving.
  - **EarlyStopping:** Stops training if validation accuracy does not improve for 10 epochs.

By combining CNN's feature extraction with LSTM's sequential learning, the model improves its understanding of facial expressions, making it more effective for emotion recognition. [8]

4.3.3 DenseNet

4.3.3.1 Advantages and Unique Features

DenseNet121 is used in this FER model because it is a powerful DL architecture known for its efficient feature reuse. Unlike traditional deep networks, where information is passed layer by layer, DenseNet connects each layer to every other layer in a feed-forward manner. This design helps in reducing the problem of vanishing gradients and allows the model to learn better with fewer parameters. DenseNet is also computationally efficient because it reduces redundant features, making it a good choice for complex image classification tasks like emotion detection. Another key advantage is that DenseNet requires fewer parameters compared to other deep networks, leading to faster training and lower memory usage. [9]

4.3.3.2 DenseNet Architecture

- **Base Model:** DenseNet121 (pre-trained on ImageNet) is used as a feature extractor.
- **Global Average Pooling Layer:** Reduces the feature map size and converts it into a vector for classification.
- **Fully Connected Layers:**
  - **512 neurons (ReLU activation) + Dropout (0.5):** Captures deep features and prevents overfitting.
  - **256 neurons (ReLU activation) + Dropout (0.3**): Adds another layer for better learning.

- **Softmax Output Layer:** Classifies emotions into different categories.
  - **Frozen Pre-trained Layers:** The lower layers of DenseNet121 are not trained again to retain learned features.
- **Optimizer:** Adam (learning rate = 0.0001) is used for stable and efficient learning [9].

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this study, we compared three DL models: CNN, CNN+LSTM (Proposed Model), and DenseNet-121 for FER.

5.1 Evaluation of the Performance of Models based on Recall, F1 Score, AUC Score, and Precision

Model	Recall score	F1 score	Auc score	Precision score
CNN	0.64	0.61	0.91	0.65
CNN+LSTM (Proposed Model)	0.71	0.66	0.91	0.66
DenseNet-121	0.53	0.51	0.84	0.52

Table 1: Evaluation of Models' (CNN, CNN+LSTM and DenseNet-121) Performance based on Recall, F1 Score, AUC Score and Precision

The CNN+LSTM Proposed Model performed the best in terms of recall and F1 score, indicating that it captures more meaningful patterns in facial emotions. CNN performed slightly lower, but still achieved a good AUC score. DenseNet-121, however, had the lowest performance, likely due to the limited number of training epochs (30) and hardware constraints.

5.2 Accuracy and Loss Graphs

Below are the training and validation accuracy and loss graphs for CNN+LSTM and DenseNet-121. These graphs help visualize how each model learned over the epochs.

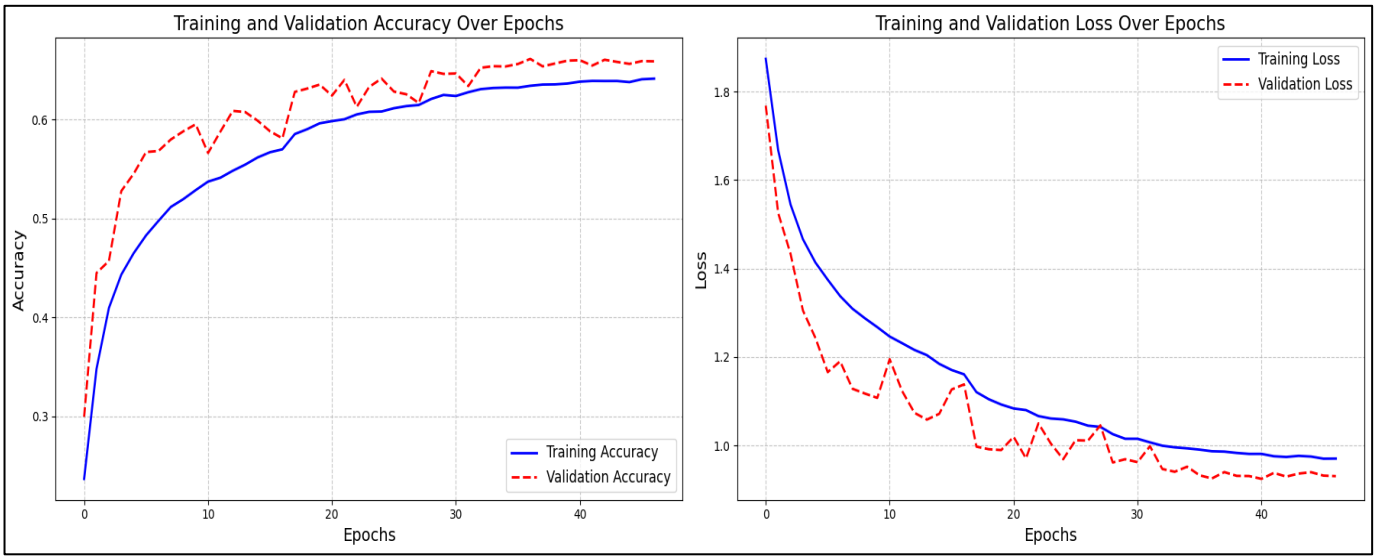

Figure 3: CNN Accuracy and Loss Graphs

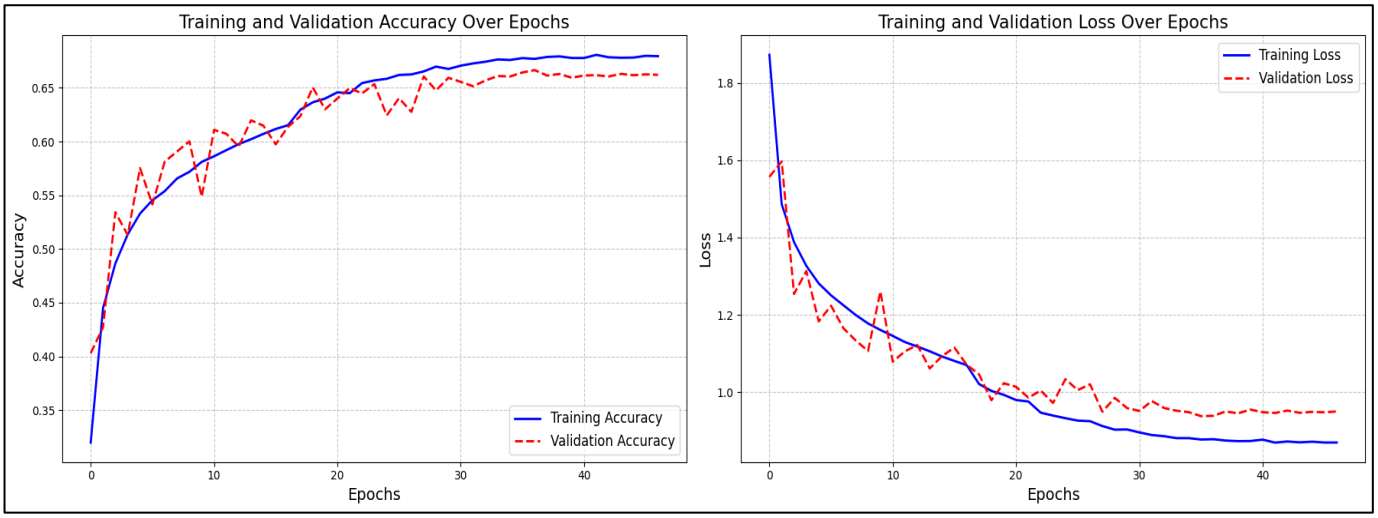

Figure 4: CNN+LSTM Accuracy and Loss Graphs

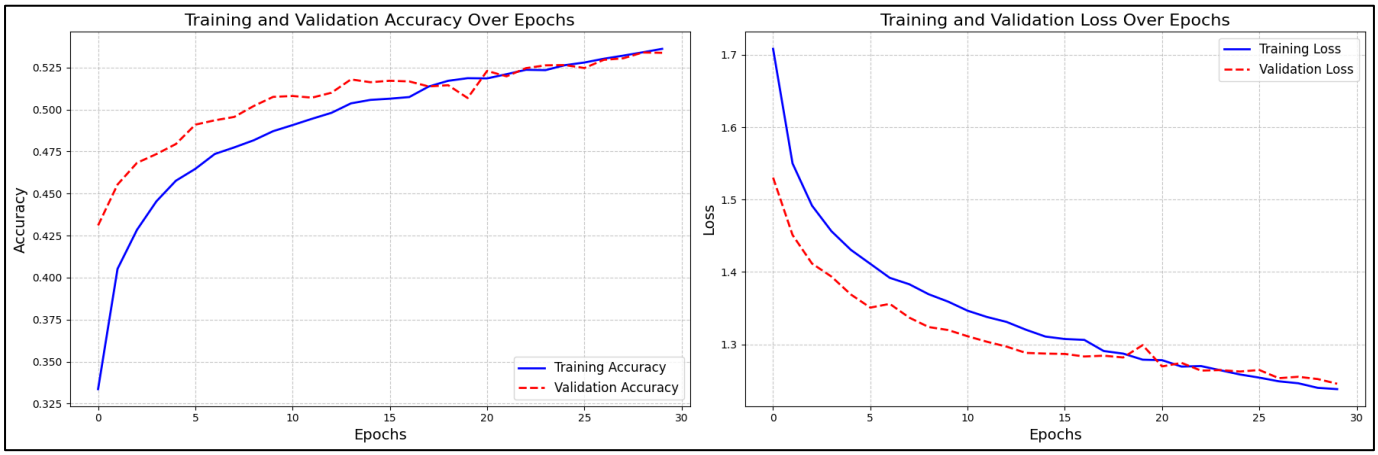

Figure 5: DenseNet-121 Accuracy and Loss Graphs

Observation:

- CNN+LSTM presented the most even learning curve which indicated balanced learning.
- The accuracy curve of DenseNet121 remained lower because the model either overfit the data or failed to extract enough features.

5.3 Training vs. Test Accuracy Bar Graph

The bar graph helps us understand generalization by comparing training and test accuracy of all three models.

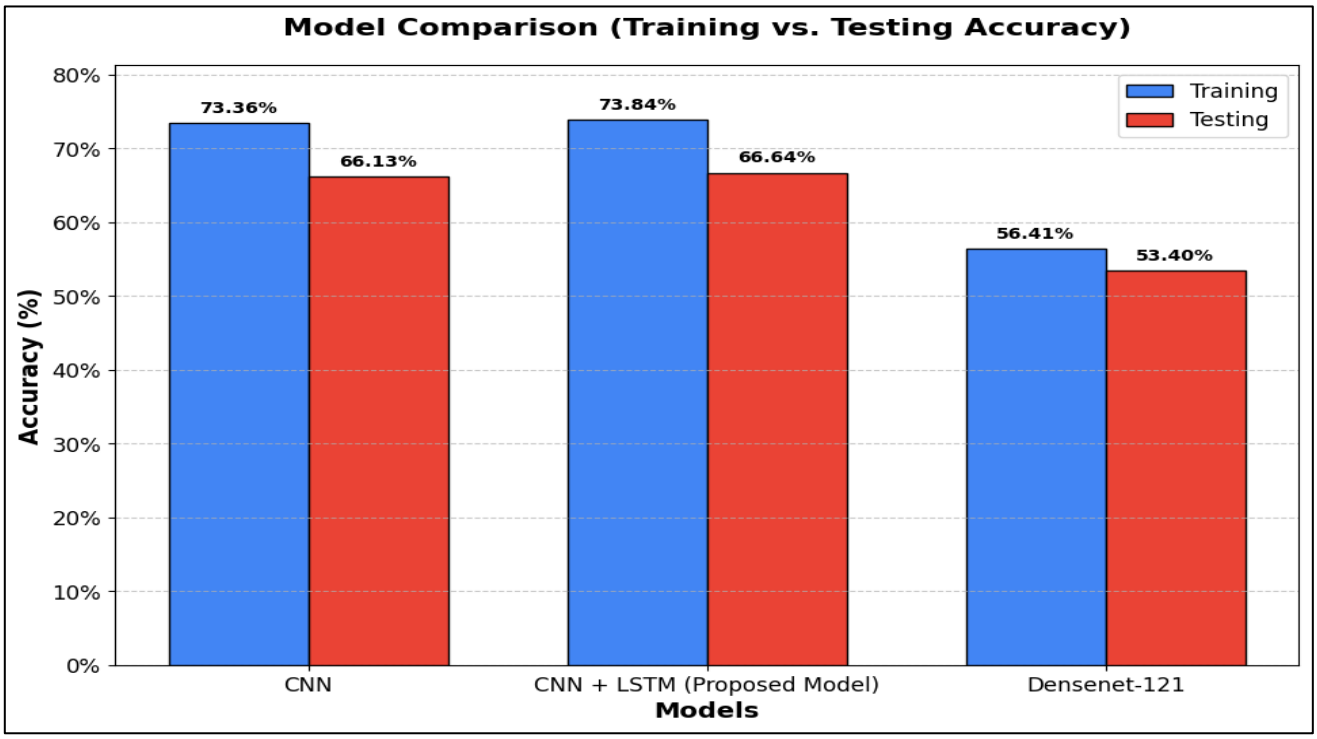

Figure 6: Model Comparison (Training vs. Testing Accuracy)

The CNN+LSTM model demonstrates the optimal trade-off between training and test accuracy while DenseNet-121 fails to generalize well because of hardware limitations and a shorter training period.

5.4 Strengths and Limitations of Each Model

The FER models present different benefits together with various difficulties in their operation.

5.4.1 CNN

The CNN model is simple and efficient, making it a good choice for systems with limited computing power. The model achieves excellent results in extracting vital image features which leads to good AUC score and precision results. The main disadvantage of this model is its inability to analyze the sequence of facial expressions throughout time.

5.4.2 The CNN+LSTM Proposed Model

The proposed model delivers the highest recall and F1 score among all three models. The model achieves improved emotion recognition through its combination of CNN for spatial feature extraction and LSTM for temporal pattern learning. The model demonstrates its reliability through its high AUC score.

5.4.3 DenseNet-121

The DenseNet-121 model is a powerful DL architecture designed to reuse features efficiently. The model extracts deep meaningful features from images which makes it suitable for various classification tasks. The study did not achieve the anticipated performance level from this model. The model failed to reach its full potential because training was restricted to 30 epochs to avoid overfitting and decrease computational time. The model generated inferior recall and F1 scores than the CNN+LSTM model. The system requires substantial computing power which makes it inefficient for hardware-constrained systems.

6. CONCLUSION

The research investigated three DL models for FER which included CNN, CNN+LSTM and DenseNet-121. The different models demonstrated distinct advantages and disadvantages in their performance. The CNN model achieved good results in precision and AUC scores yet failed to detect sequential patterns in emotions. The CNN+LSTM model produced the best results because it combined spatial and temporal features to achieve the highest recall and F1 score. The DenseNet-121 deep network showed poor performance because of restricted training epochs combined with hardware limitations.

Among these models, the CNN+LSTM Proposed Model proved to be the best. The model achieved superior emotion recognition results because it processed facial features together with time-dependent patterns. This model stands as an excellent selection for systems that need precise emotion detection.

Multiple potential improvements exist for future development. Vision Transformers (ViTs) represent transformer-based models that could enhance accuracy through their ability to detect

global image dependencies. The models will achieve better generalization for face types and lighting conditions and expressions when trained on expanded diverse datasets. Real-time application optimization would enable the model to serve mental health monitoring and human-computer interaction systems as well as security systems. Future FER systems will become more reliable and useful through improvements in both accuracy and efficiency.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1, no. 2. Cambridge, MA: MIT Press, 2016.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539

[3] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2020. doi: 10.1109/TAFFC.2020.2981446

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90

[5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017. doi: 10.1109/CVPR.2017.243

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735

[7] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 279–283, Oct. 2016. doi: 10.1145/2993148.2993165

[8] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018. doi: 10.1109/TIP.2018.2886767

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, 2012. doi: 10.1109/MMUL.2012.26

[10] K. Wang, C. Liu, and S. Shen, "Geometric calibration for cameras with inconsistent imaging capabilities," *Sensors*, vol. 22, no. 7, p. 2739, 2022. doi: 10.3390/s22072739

[11] Kaggle, "FER-2013," *Kaggle*, [Online]. Available: https://www.kaggle.com/datasets/msambare/fer2013. [Accessed: Apr. 28, 2025].

[12] Kaggle, "RAF-DB DATABASET," *Kaggle*, [Online]. Available: https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset. [Accessed: Apr. 28, 2025].

[13] J. Yun et al., "Deep radiomics-based survival prediction in patients with chronic obstructive pulmonary disease," *Scientific Reports*, vol. 11, no. 1, p. 15144, 2021. doi: 10.1038/s41598-021-94535-4

[14] Adobe Stock, "Facial Expression," *Adobe Stock*, [Online]. Available: https://stock.adobe.com/search?k=facial+expression. [Accessed: Apr. 28, 2025].

[15] iStock, "Facial Expression," *iStock*, [Online]. Available: https://www.istockphoto.com/photos/facial-expression. [Accessed: Apr. 28, 2025].

APPENDIX

Acronyms

AI	Artificial Intelligence
AUC	Area Under the Curve
CNN	Convolutional Neural Networks
DL	Deep Learning
FER	Facial Emotion Recognition
HCI	Human-Computer Interaction
HOG	Histogram of Oriented Gradients
LBP	Local Binary Patterns
LSTM	Long Short-Term Memory
ML	Machine Learning
RAF-DB	Real-world Affective Faces Database
ReLU	Rectified Liner Unit
RNN	Recurrent Neural Networks
SMOTE	Synthetic Minority Over-sampling Technique