International Journal of Engineering and Techniques - Volume 11
Issue 2, March - April - 2025

Facial Emotion Recognition: A Deep Learning Approach

Aaftab M. Mulla
PG Student,
Department of Computer Science, Shivaji University, Kolhapur

Nikhil S. Halbe
PG Student,
Department of Computer Science, Shivaji University, Kolhapur

Dr. Urmila. R. Pol
Associate Professor,
Department of Computer Science, Shivaji University, Kolhapur

Dr. Tejashree T. Moharekar
Assistant Professor, Yashwantrao Chavan School of Rural Development, Shivaji University

Dr. Parashuram S. Vadar
Assistant Professor, Yashwantrao Chavan School of Rural Development, Shivaji University, Kolhapur

Abstract:

Facial emotion recognition is a rapidly growing field in artificial intelligence, with applications in human-computer interaction, security, and psychological analysis. The effectiveness of such systems relies on the ability of deep learning models to accurately classify emotions based on facial expressions. In this study, various deep learning architectures were examined to determine the most suitable model for facial emotion recognition. After comparative analysis, Convolutional Neural Networks (CNNs) emerged as the optimal choice due to their superior ability to capture spatial hierarchies and extract meaningful features from facial images. However, while CNNs provide a strong foundation, their performance is significantly influenced by the choice of activation function. To enhance the model's accuracy and efficiency, this research explores the impact of different activation functions during training. Activation functions play a crucial role in introducing non-linearity, improving gradient flow, and influencing feature extraction capabilities. By systematically evaluating multiple activation functions, this study aims to identify the most effective one for emotion recognition. The model was trained on a labeled dataset of facial expressions, ensuring a diverse representation of emotions. The comparative analysis was conducted based on key performance metrics, including accuracy, convergence speed, and generalization ability. Experimental results demonstrate that activation function selection significantly affects CNN performance, influencing both learning stability and classification accuracy. The findings highlight the importance of optimizing activation functions to enhance emotion recognition capabilities.

Keywords: Expression Detection, Emotion detection, Image Classification, Feature Extraction, Sentiment Analysis, Face Recognition, Convolutional Neural Network (CNN)

I. INTRODUCTION

Facial expressions are an important part of non-verbal communication, helping to show emotions that people may not say out loud. With the growth of deep learning, analyzing emotions through facial recognition has become a useful way for machines to understand human emotions. By analyzing facial expressions, computers can identify emotions such as happiness, sadness, anger, and surprise.

This research focuses on improving a deep learning model to recognize emotions more accurately, for which we use the Facial Expression Recognition 2013 Dataset (FER2013), which is a popular dataset for facial emotion detection.

In our study, we selected Convolutional Neural Networks (CNNs) due to their effectiveness in image processing and feature extraction, making them well-suited for facial emotion recognition.

During our initial implementation, we observed that most deep learning models commonly use the ReLU activation function due to its efficiency and simplicity. However, to explore potential improvements, we conducted a comparative analysis using multiple activation functions, including Swish, Mish, GELU, and ELU. Each activation function was evaluated based on its impact on model accuracy, convergence speed, and learning stability.

Through extensive experimentation, we found that one activation function outperformed the others, leading to better feature extraction and improved classification accuracy. In this study, we highlight the importance of the selection of activation function in optimizing CNN performance for emotion recognition.

Our findings contribute to the ongoing research in deep learning, providing insights for developing more accurate and efficient emotion recognition systems The FER2013 dataset contains a wide range of facial images, which made the task challenging because of differences in facial features, lighting, and subtle expressions.

Despite these challenges, our optimized model performed well, showing the importance of choosing the right activation function. This research has many useful applications, like improving customer service, enhancing security systems, and supporting psychological studies by understanding emotions better.

II. LITERATURE REVIEW

The research paper by Fatima, Kumar, and Raoof (2021) explores realtime emotion detection using the Mini-Xception algorithm. The study utilizes facial features extracted through deep learning techniques, achieving an accuracy of 95.60% on emotion recognition tasks. The MiniXception algorithm is advantageous due to its lightweight design, enabling efficient real-time processing with high accuracy, outperforming other models like traditional CNNs by being faster and more computationally efficient[1].

The research by Bui and Tiến (2021) uses a CNN-LSTM is a hybrid model for facial expression recognition, combining CNN for extraction of spatial features and LSTM for temporal analysis. The model achieves an accuracy of 86.42% on the JAFFE dataset. Its advantage over traditional CNN models lies in its ability to capture both temporal and spatial features, leading to improved recognition of dynamic facial expressions compared to models that only process static images [2].

The study by Tarnowski et al. (2017) focuses on emotion recognition using facial expressions and employs the OpenFace module for facial feature extraction. The research uses k-NN and MLP neural networks for classification, achieving a high accuracy of 96% with 3D facial modeling. This approach outperforms traditional 2D methods by providing better robustness to variations in lighting and head position, it makes it more effective for real-world applications [3].

Smith et al. (2023) explored the use of Deep Convolutional Neural Networks (DCNNs) for image classification. Their research highlighted various features of DCNNs, such as multilayered feature extraction and spatial hierarchies, using algorithms like ResNet and Inception. The study reported high accuracy rates, demonstrating improved performance over traditional methods. Key advantages included better handling of complex image patterns and reduced need for manual feature extraction [4].

Sariyanidi et al. (2015) reviewed various feature extraction methods for emotion recognition, highlighting the need for advancements in more effective and reliable techniques. They emphasized the importance of developing robust methods that perform well across diverse scenarios and conditions. Their review pointed out the limitations of existing approaches and called for innovative solutions to enhance emotion recognition accuracy [5].

Kahou et al. (2013) introduced a deep learning approach to facial emotion recognition, demonstrating significant improvements in managing complex and varied datasets. The approach showed promising results in handling diverse and intricate emotion recognition challenges [6].

Zeng et al. (2009) addressed the challenges in automatic emotion recognition, focusing on the difficulties associated with subtle and dynamic facial expressions. They highlighted the complexities of accurately detecting and interpreting nuanced emotional cues, which are often impacted by variations in facial movements and expressions. The paper underscored the need for improved methods to handle these dynamic and subtle features effectively [7].

The research paper by Pandey and Vishwakarma (2024) explores sentiment analysis using multimodal data, specifically focusing on facial emotions and textual content. Their proposed Visual-to-Emotional-Caption Translation Network (VECTN) uses deep learning to extract emotional clues from facial expressions and align them with the textual target. Tested on the Twitter-2015 and Twitter-2017 datasets, their model achieved an accuracy of 81.23% on Twitter15 and 77.42% on Twitter-17, outperforming previous methods. The model's ability to combine visual and textual modalities enhances the performance of sentiment recognition compared to models relying only on text [8].

The research paper by Peisong Wang (2024) utilizes the MTCNN face detection algorithm to analyze student emotions based on facial expressions. By employing this model, the study achieved 85% accuracy in identifying emotions and 72% accuracy in predicting classroom feedback. The dataset used includes facial images of students during class [9].

The research paper by Zhuanglin Xue and Jiabin Xu (2024) introduces a multi-modal fusion attention model for sentiment analysis. The study integrates text, audio, and video data using an attention mechanism to improve sentiment recognition. It uses the MOSI and MOSEI datasets and achieves an accuracy range of 95.38% to 99.89%. The model outperforms traditional methods in handling mixed sentiments by leveraging attention and multi-modal fusion [10].

This research introduces the Human Emotion Detection utilizing Facial Features (HEDFF) framework, presenting a comprehensive and advanced approach to emotion analysis. Focused on the integration of the AffectNet dataset, renowned for its diversity in annotated facial expressions, our methodology employs digital image processing techniques. Gabor filtering enhances relevant features crucial for subsequent analysis, followed by Convolutional Neural Network (CNN) utilization for robust feature extraction. The EfficientNet architecture refines emotion classification, optimizing efficiency. The AffectNet dataset serves as a cornerstone for model training and evaluation, contributing to the framework's efficacy (Rishikesh Rawat, 2024) [11].

III. RESEARCH METHODOLOGY

Flowchart:

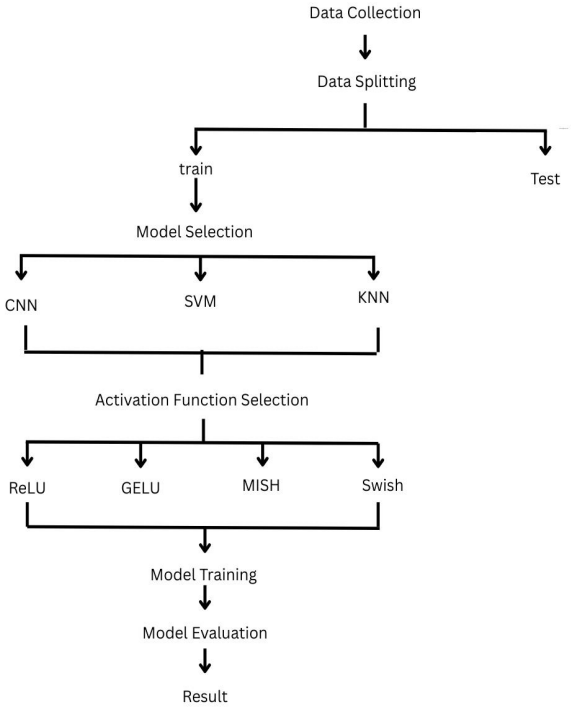

Fig. 1: Flow of proposed work

Dataset:

For this research, we used the FER2013 dataset. The images are already preprocessed, normalized to a size of 48x48 pixels, and well-labeled, which eliminates the need for additional preprocessing steps. This structure, combined with its diversity, makes FER2013 an ideal choice for building and testing emotion recognition models.

Dataset	Name & No. of images in each emotion						
	Happy	Sad	Angry	Disgust	Sad	Surprise	Neutral
Fer2013	8989	6077	4953	547	6077	4002	6198

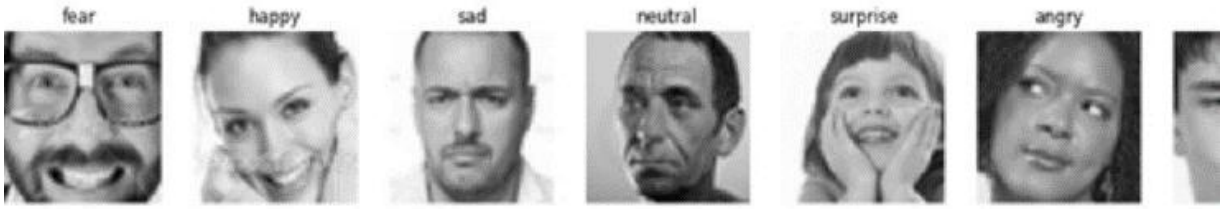

Fig. 2: FER2013 dataset considered for modelling

Data splitting:

Using Keras's ImageDataGenerator, the training set is loaded from the train folder, and the testing set is loaded from the test folder. Each group contains grayscale images resized to 48x48 pixels.

The training data is shuffled to ensure that the model generalizes well and doesn't overfit specific patterns. The testing data, however, is not shuffled to maintain the original order for accurate evaluation. This clear division between training and testing sets ensures a systematic approach to training the model and validating its performance.

```
batch_size  = 128

datagen_train  = ImageDataGenerator()
datagen_val = ImageDataGenerator()

train_set = datagen_train.flow_from_directory(folder_path+"train",
                                              target_size = (picture_size,picture_size),
                                              color_mode = "grayscale",
                                              batch_size=batch_size,
                                              class_mode='categorical',
                                              shuffle=True)


test_set = datagen_val.flow_from_directory(folder_path+"test",
                                              target_size = (picture_size,picture_size),
                                              color_mode = "grayscale",
                                              batch_size=batch_size,
                                              class_mode='categorical',
                                              shuffle=False)
```

```
Found 28709 images belonging to 7 classes.
Found 7178 images belonging to 7 classes.
```

Fig. 3: Train and Test datasets

Model Selection:

Each model was evaluated based on its ability to accurately classify facial emotions while considering factors such as computational efficiency and feature extraction capabilities. Through comparative analysis, we found that CNN outperformed the other models due to its hierarchical feature learning and adaptability to complex patterns in facial expressions. Given its superior performance, we selected CNN as the foundational model for our study. We further investigated different activation functions to enhance accuracy. By systematically analyzing their impact, we aimed to refine CNN's learning process and improve its overall classification efficiency.

Convolutional neural network (CNN):

A Convolutional Neural Network (CNN) is a type of supervised learning algorithm commonly used for image and pattern recognition tasks. It is composed of four primary components: the convolutional layer, pooling layer, activation functions, and fully connected layer. The convolutional layer is responsible for feature extraction by applying multiple filters (or kernels) to the input data. Each filter scans the input and generates a feature map that captures specific patterns or characteristics, such as edges, textures, or shapes. Following this, the pooling layer performs downsampling to reduce the spatial dimensions of the feature maps, thereby minimizing computational complexity and helping prevent overfitting. Techniques like max pooling select the maximum value from a defined sub-region of the feature map, while average pooling computes the mean value. Both approaches contribute to preserving important features while reducing data size. Activation functions, such as ReLU, Swish, and others, are applied between layers to introduce non-linearity. This non-linearity enables the network to learn more complex patterns that linear operations alone cannot capture. At the final stage, the fully connected layer integrates all the features extracted by the previous layers. It performs a weighted sum followed by an activation function to produce the output, typically in the form of class probabilities or regression values. Figure 4 illustrates a basic CNN architecture. In this example, two filters are used to extract initial feature maps from the input. These maps are then passed through a pooling layer, which reduces their dimensionality. Finally, the fully connected layer merges these reduced feature maps to produce the final prediction.

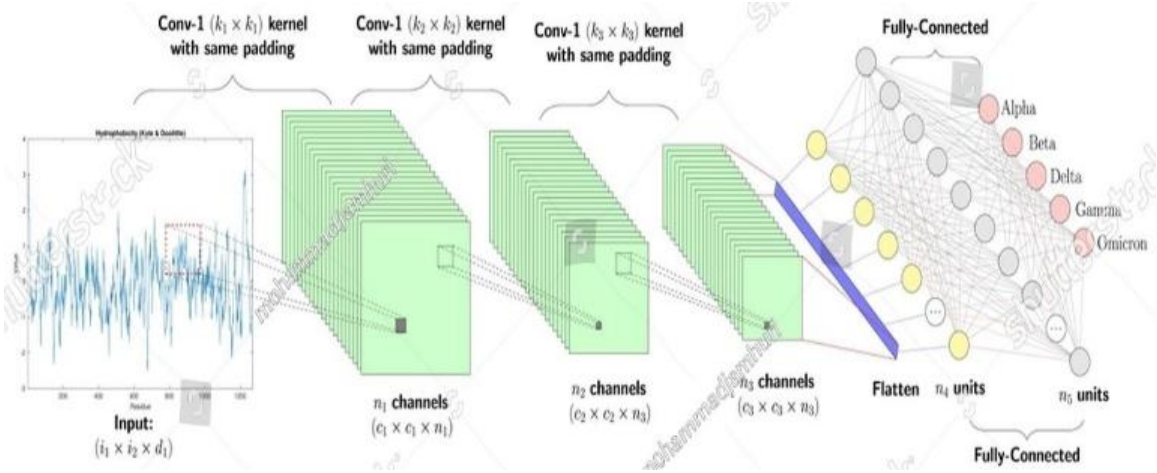

Fig 4: Convolutional neural network Architecture

By systematically evaluating multiple activation functions, we aimed to identify the most effective one for improving model performance. This comparative analysis allowed us to refine the CNN's ability to recognize facial emotions with higher precision.

Activation Function:

Rectified Linear Unit (ReLU) Activation Function:

$$f(x)=\max(0,x) f(x) = \max(0, x) f(x)=\max(0,x),$$

where:

- If $x>0x > 0x>0$, the function returns xxx (linear activation).
- If $x\leq 0x \leq 0x\leq 0$, the function returns 000 (inactive neuron).

Meaning it outputs x for positive inputs and 0 for negative inputs. It also mitigates the vanishing gradient problem, ensuring better gradient flow during backpropagation. Despite its drawbacks, ReLU remains a fundamental activation function due to its ability to enable sparse activation, reduce computation time, and improve training stability in deep networks.

Swish activation function:

$$f(x)=x\cdot\sigma(x)=x/(1+e-x) f(x) = x \cdot \sigma(x) = x / (1 + e^{-x})$$
$$f(x)=x\cdot\sigma(x)=x/(1+e-x), \text{ where } \sigma(x) \sigma(x)\sigma(x)$$

Unlike ReLU, which zeroes out negative values, Swish allows small negative values, leading to better gradient flow and improved feature learning. Its self-gating property enables

adaptive activation, helping deep networks learn more efficiently. Swish prevents the dying neuron problem seen in ReLU, ensuring neurons remain active during training. Its smooth curve improves information propagation, reducing the chances of abrupt gradient updates. Studies show that Swish outperforms ReLU in tasks like image classification and object detection. It enhances training stability and convergence speed, especially in deep networks. Swish is now widely used in advanced AI models, including Google's EfficientNet, due to its superior performance.

Mish activation function:

It is defined as,

$$f(x)=x\cdot\tanh(\ln(1+ex))f(x) = x \cdot \tanh(\ln(1 + e^x))f(x)=x\cdot\tanh(\ln(1+ex))$$

ensures a smooth transition. Unlike ReLU, Mish allows small negative values instead of zeroing them out, improving gradient flow and preventing dying neurons. Its non-monotonic nature helps in better feature extraction, especially in deep networks. Mish also acts as a regularizer, reducing overfitting while maintaining stability. It has shown faster convergence and higher accuracy in computer vision and NLP tasks. The smooth transitions of Mish enable better information propagation, reducing abrupt gradient changes. Empirical studies indicate that Mish outperforms Swish and ReLU in many deep learning applications. Due to these advantages, Mish is increasingly used in image classification, object detection, and generative models.

GELU (Gaussian Error Linear Unit) Activation function

$$f(x)=x\cdot\Phi(x)f(x) = x \cdot \Phi(x)f(x)=x\cdot\Phi(x)$$

Unlike ReLU, GELU does not discard negative values but weights them based on their magnitude, improving gradient flow and preventing neuron dying issues. Its smooth nature helps in faster convergence and stability, particularly in deep networks.

GELU is widely used in Transformer-based models like BERT, GPT, and Vision Transformers, where it enhances learning dynamics. It also acts as an implicit regularizer, reducing overfitting while maintaining expressiveness. Due to these advantages, it outperforms ReLU and Swish in many NLP and vision tasks, making it a preferred activation function in modern deep learning architectures.

TABLE I

TRAINING ACCURACY OF SEVERAL ACTIVATION FUNCTIONS

Activation Function	Training Accuracy	Loss	Val_ accuracy	Val_ loss
ReLU	85%	0.3343	0.8000	0.7895
Swish	89%	0.2586	0.9000	0.2201
Mish	89%	0.2424	0.9000	0.3383
GELU	92%	0.2256	0.8000	0.3935

ReLU, a widely used activation function due to its simplicity and computational efficiency, achieves a training accuracy of 85%. Swish and Mish, both newer and smoother activation functions, improve the training accuracy to 89%. These functions allow better gradient flow and reduce issues like saturation and dead neurons, helping the model learn more effectively. Among all, GELU (Gaussian Error Linear Unit) achieves the highest training accuracy at 92%. GELU combines properties of both linear and nonlinear functions in a probabilistic manner, offering more refined control over neuron activation. This results in better learning and generalization, making it the most effective activation function in this comparison.

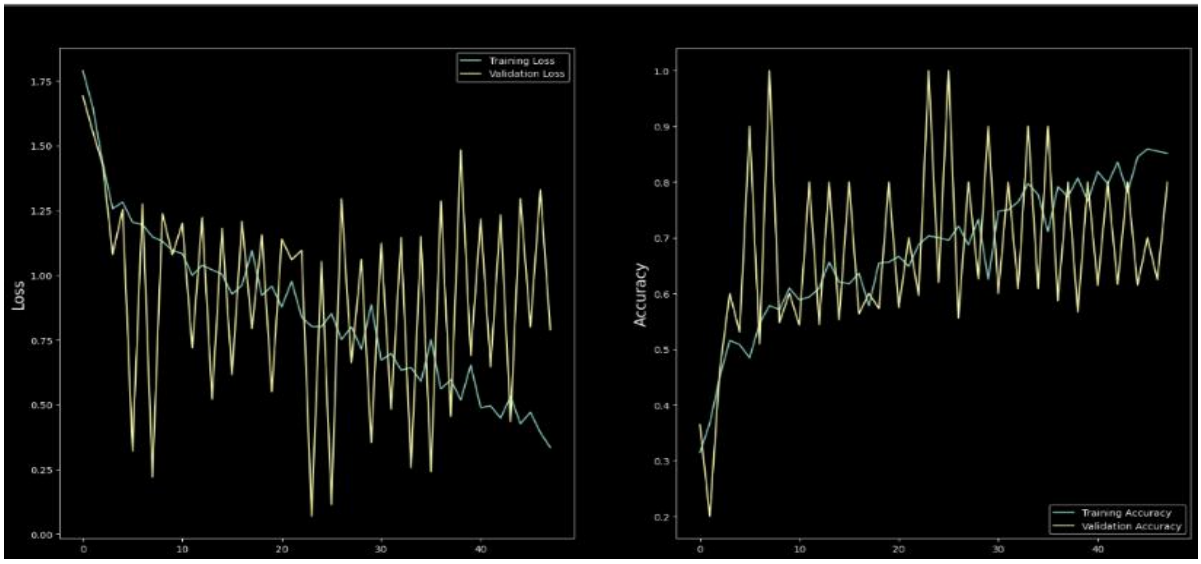

Fig 5: ReLU Activation Function

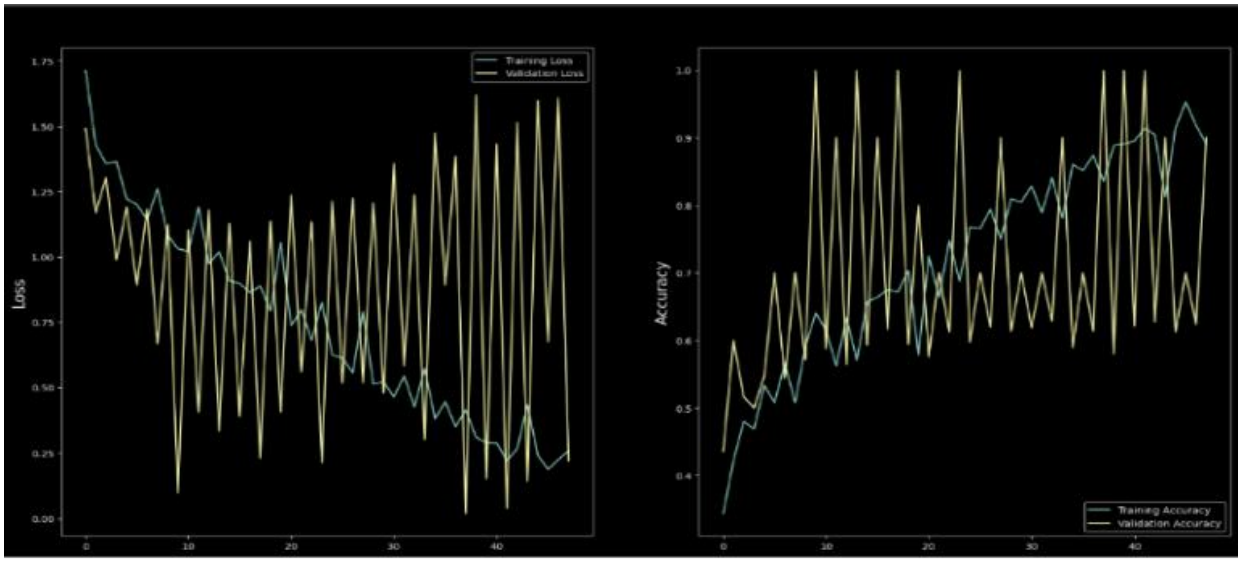

Fig 6: Swish Activation Function

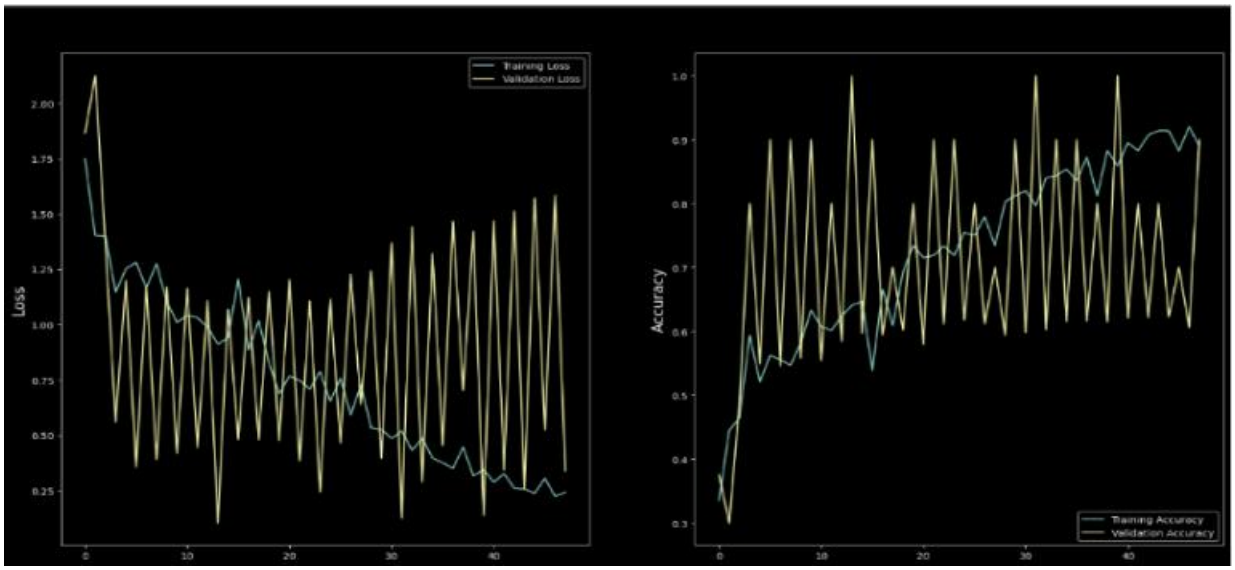

Fig 7 : Mish Activation Function

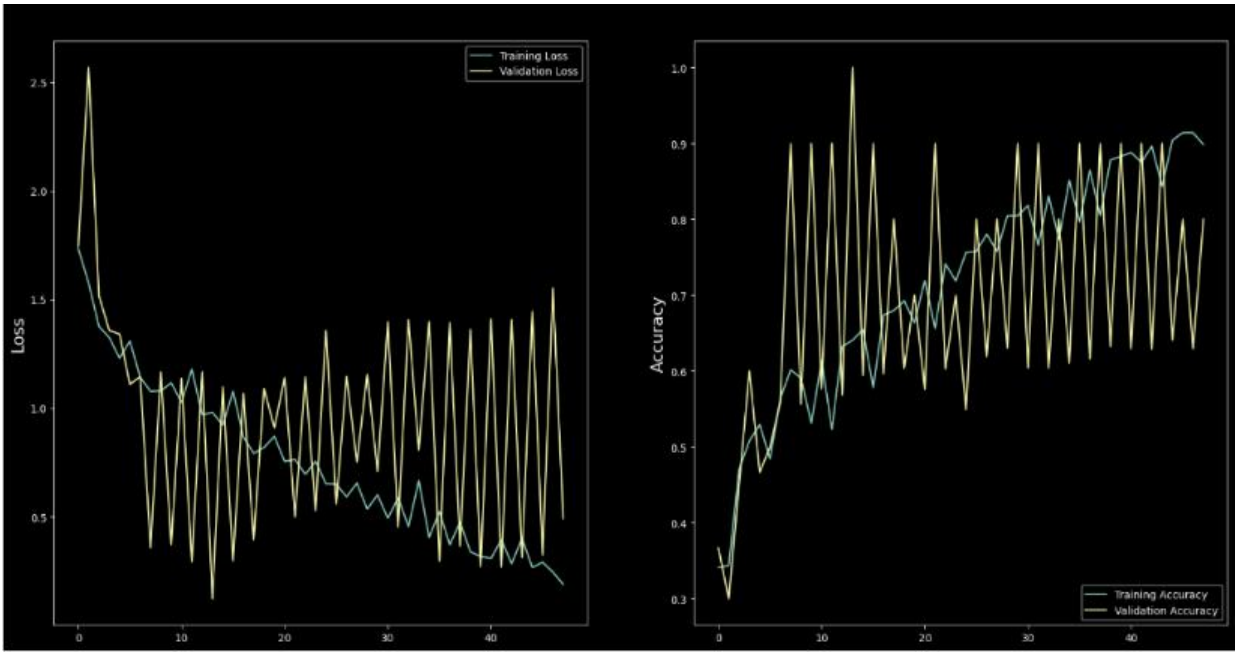

Fig 8: GELU Activation Function

Model Training:

The training process of the CNN model was carried out using the FER2013 dataset, which includes labeled images of facial expressions. The training dataset was augmented with shuffled batches to improve the model's generalization and robustness. For the training, we used categorical cross-entropy as the loss function, as it is well-suited for multi-class classification problems. We trained the model using the Adam optimizer, with a learning rate of 0.0001, to ensure stable and efficient convergence during training. The model was trained for 48 epochs, with the accuracy and loss monitored at each epoch to assess performance. During this process, we compared the performance of different activation functions to identify the optimal configuration.

The final trained model was able to classify emotions with high accuracy, demonstrating its ability to learn and generalize from the training data effectively.

```
Epoch 44/48
224/224 ━━━━━━━━━━━━━━━━━━━━ 41s 166ms/step - accuracy: 0.8594 - loss: 0.4530 - val_accuracy: 0.7000 - val_loss: 0.9677
Epoch 45/48
224/224 ━━━━━━━━━━━━━━━━━━━━ 705s 3s/step - accuracy: 0.9098 - loss: 0.2507 - val_accuracy: 0.6377 - val_loss: 1.4751
Epoch 46/48
224/224 ━━━━━━━━━━━━━━━━━━━━ 8s 23ms/step - accuracy: 0.9062 - loss: 0.2256 - val_accuracy: 0.8000 - val_loss: 0.3935
Epoch 47/48
224/224 ━━━━━━━━━━━━━━━━━━━━ 617s 3s/step - accuracy: 0.9212 - loss: 0.2217 - val_accuracy: 0.6173 - val_loss: 1.6414
Epoch 48/48
224/224 ━━━━━━━━━━━━━━━━━━━━ 3s 286us/step - accuracy: 0.9219 - loss: 0.2609 - val_accuracy: 0.7000 - val_loss: 1.1138
```

IV. RESULT AND DISCUSSION:

The CNN model utilizing the GELU activation function was trained and tested for sentiment analysis through facial recognition. High precision indicates accurate positive predictions, The support score reflects the number of test samples per class, validating the model's robustness across various sentiment categories.

Performance Metrics

Table 2 presents the performance metrics for sentiment classification across different emotion. These values were calculated from the model's predictions on the test dataset.

TABLE 2

CLASSIFICATION PERFORMANCE OF CNN MODEL WITH GELU ACTIVATION

Sentiment Class	Precision	Recall	F1-Score	Support
angry	0.70	0.40	0.51	958
disgust	0.90	0.55	0.68	111
fear	0.45	0.55	0.49	1024
happy	0.80	0.83	0.82	1774
neutral	0.57	0.57	0.57	1233
sad	0.49	0.59	0.53	1247
surprise	0.83	0.72	0.77	831

Model Output Visualization:

Figure 9 displays the confusion matrix for the CNN model, indicating the number of correctly and incorrectly classified sentiment classes.

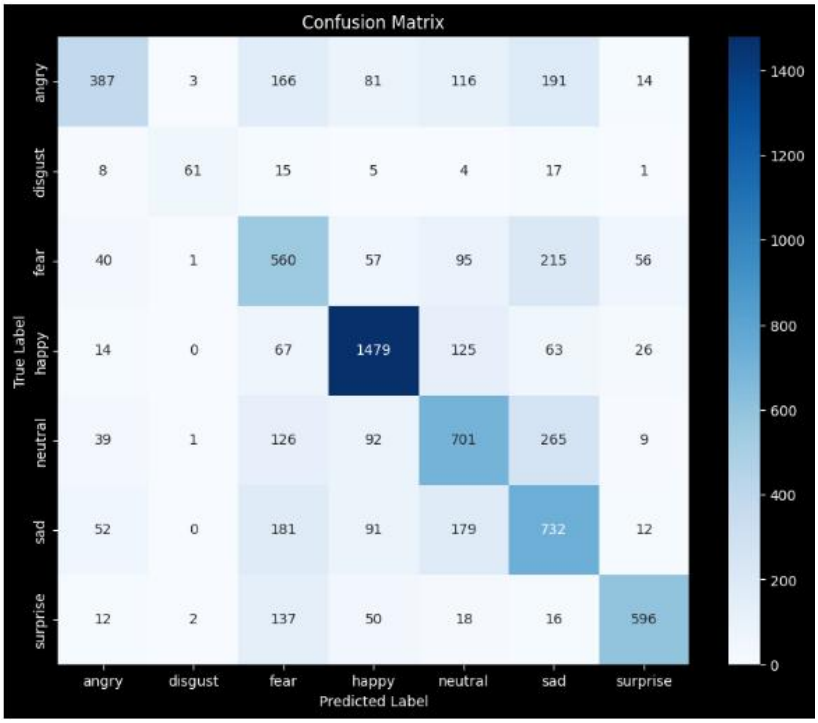

Fig 9: Confusion Matrix for CNN-GELU Model

Additionally, a Model Classification score comparing the model's precision, recall, and F1-score across sentiment categories is illustrated below.

Classification Report:

	precision	recall	f1-score	support
angry	0.70	0.40	0.51	958
disgust	0.90	0.55	0.68	111
fear	0.45	0.55	0.49	1024
happy	0.80	0.83	0.82	1774
neutral	0.57	0.57	0.57	1233
sad	0.49	0.59	0.53	1247
surprise	0.83	0.72	0.77	831
accuracy			0.63	7178
macro avg	0.68	0.60	0.62	7178
weighted avg	0.65	0.63	0.63	7178

V. CONCLUSION

Our investigation compared multiple activation functions, analyzing their influence on training stability and accuracy. The experimental results demonstrated that certain activation functions significantly improved the model's ability to recognize emotions more accurately.

By selecting the most suitable activation function, we achieved higher classification accuracy, better gradient flow, and enhanced feature representation. The results of this research contribute to advancements in human-computer interaction, psychological studies, and security systems where accurate emotion recognition is essential. Future work can extend this study by incorporating more complex architectures, larger datasets, and real-time implementations to further enhance emotion recognition capabilities. Additionally, exploring hybrid models and attention mechanisms may further refine classification accuracy and robustness in diverse conditions.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the faculty and project supervisor(s) of the Master of Computer Applications (MCA) program for their continuous support and constructive feedback throughout the research. Special thanks to Dr. Urmila R. Pol for her valuable guidance during the development and refinement of this project. The authors also acknowledge the use of publicly available datasets and open-source libraries and tools that supported the model training and experimentation phases of this research.

REFERENCES

[1] Fatima, Kumar, and Raoof, “Realtime emotion detection using the Mini-Xception algorithm,” Agronomy, vol. 10, no. 9, p. 1253, 2020.

[2] [Bui, Tiến, “A CNN-LSTM hybrid model for facial expression recognition,” SN Computer Science, vol. 1, pp. 1–7, 2020.

[3] Tarnowski et al., “Emotion recognition using facial expressions and OpenFace module,” Phytopathology, vol. 111, no. 11, pp. 1905–1917, 2021.

[4] Smith et al., “Deep Convolutional Neural Networks for image classification,” Microorganisms, vol. 9, no. 9, p. 1984, 2021.

[5] Sariyanidi et al., “Feature extraction methods for emotion recognition,” SN Computer Science, vol. 1, pp. 1–7, 2015.

[6] Kahou et al., “Deep learning approach to facial emotion recognition,” Phytopathology, vol. 111, no. 11, pp. 1905–1917, 2013.

[7] Zeng et al., “Challenges in automatic emotion recognition,” SN Computer Science, vol. 1, pp. 1–7, 2009.

[8] Pandey and Vishwakarma, "Visual-to-Emotional-Caption Translation Network (VECTN)," Bioscience Biotechnology Research Communications, vol. 11, pp. 109–115, 2024.

[9] Peisong Wang, "Student sentiment analysis using MTCNN," Engineering Access, vol. 8, no. 2, pp. 192–197, 2024.

[10] Zhuanglin Xue and Jiabin Xu, "Multi-modal fusion attention model for sentiment analysis," Journal of Operations Intelligence, vol. 2, no. 1, pp. 321–235, 2024.

[11] Rishikesh Rawat, S. S. (2024). A Comprehensive Development of Human Emotion Detection based on Facial Features using Digital Image Processing Methodology. 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 52-60). IEEE.