International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 03 | Mar 2025 www.irjet.net p-ISSN: 2395-0072

Optimizing Facial Expression Recognition with Deep Hybrid Neural Networks

Satyam Kumar[1] and Purusharth Agarwal[2]

[1]*Student, Department of AIML, Manipal University Jaipur, Rajasthan, India*
[2]*Student, Department of AIML, Manipal University Jaipur, Rajasthan, India*

Abstract - In Facial Expression Recognition (FER), the landscape has undergone a remarkable evolution in recent years, driven by significant breakthroughs in deep learning(DL), image processing, and cognitive sciences. This study endeavors to push the boundaries of FER precision and efficacy by delving deep into the intricate nuances of facial movement features in static images. The effectiveness of our proposed methodology is substantiated by compelling results shown in Table 1. Notably, Densenet121 + GRU demonstrates exceptional performance, achieving an impressive accuracy of 99.65% on the Facial Emotion Recognition Image Dataset (Table 1). Our methodology capitalizes on the fusion of DL models with RNNs, namely DenseNet121 + GRU, VGG16 + GRU and Xception + GRU, these results underscore the robustness and effectiveness of our approach in accurately discerning facial expressions from static images. Moreover, our findings shed light on the untapped potential of integrating dynamic facial movement features into FER systems. By bridging the gap between static and dynamic characteristics, our methodology holds promise in enhancing the accuracy and reliability of emotion recognition systems in real-world scenarios. In essence, this study contributes to the ongoing advancements in FER. It paves the way for future research endeavors to harness the full spectrum of facial movement features for superior emotion recognition capabilities.

Key Words: Facial Emotion Recognition, Deep Learning, VGG16, Xception, Multimodal Integration, Transfer Learning, Real-world Deployment.

1. INTRODUCTION

Facial expression recognition (FER) is a pivotal domain within AI, garnering increasing attention as it forms the linchpin for effective human-computer interaction (HCL). Emotions, conveyed through facial expressions, are fundamental components of human communication [3]. FER systems aim to decipher these expressions from static images or dynamic video sequences to discern the underlying psychological states of individuals [4]. Furthermore, in fields like counseling psychology, retail sales, social robotics, and e-learning, accurate FER holds profound implications for improving service delivery and user experience. The significance of FER is underscored by its diverse applications across multiple domains. However, the journey towards achieving robust and accurate FER systems is challenging. Traditional approaches to FER often relied on handcrafted features and shallow learning methods. Nevertheless, with the advent of DL, particularly convolutional neural networks (CNNs), significant strides have been made in improving FER accuracy. CNNs are capable of automatically extracting discriminative features from facial images, thus offering promising avenues for advancing FER technology.

The dataset utilized in this study play a pivotal role in training and evaluating our proposed FER models. The Facial Emotion Recognition Image Dataset (Kaggle) comprises 18,000 images annotated with diverse emotional states, further enriching the training data. This study explores the efficacy of several DL architectures integrated with RNNs for FER tasks. Specifically, we consider DenseNet121 + GRU, VGG16 + GRU and Xception + GRU models. These models are meticulously designed to capture intricate facial expression patterns from static images, thus enabling accurate emotion recognition. A comprehensive methodology detailing these models' architecture and training process is elucidated in Section 4. Notably, our experimentation yields compelling results, with the DenseNet121 + GRU model showcasing the highest accuracy of 99.65% on the Facial Emotion Recognition Image Dataset (Kaggle), underscoring the robustness of our proposed approach. Detailed discussions on these results are provided in Section 5, shedding light on the efficacy of our proposed FER models.

This paper comprehensively investigates DL-based FER, leveraging attention mechanisms to enhance model performance. Through extensive experimentation and analysis, we demonstrate the efficacy of our proposed approach in achieving accuracy levels on diverse FER datasets. The remainder of this paper is organized as follows: Section 2 presents related work in FER, while Section 3 delineates the motivation behind our proposed methodology. Section 4 provides a detailed exposition of our methods, encompassing dataset descriptions, data vectorization techniques, and model architectures. Section 5 presents our experimental results and analyses, followed by concluding remarks and avenues for future research in Section 6.

2. RELATED WORK

Nan et al. (2022) introduced A-MobileNet, leveraging mobile architecture for efficient FER. Their approach emphasizes the importance of lightweight models for real-time applications.[1] Wen et al. (2023) proposed a Multi-head Cross Attention Network, "Distract your attention," enhancing FER performance by capturing high-order interactions among facial features. Their method demonstrates the efficacy of attention mechanisms in FER tasks.[2] Li and Lima (2021) utilized ResNet-50 for FER, showcasing the effectiveness of CNN in capturing facial features. Their study highlights the potential of traditional DL architectures in FER.[5] Wang et al. (2020) proposed Region Attention Networks, emphasizing pose and occlusion-aware attention mechanisms for robust FER. Their approach highlights the importance of incorporating spatial attention for accurate expression recognition.[6] Mollahosseini et al. (2016) introduced DL architectures for FER, demonstrating the potential of DL in capturing facial features. Their study underscores the effectiveness of DL in FER tasks.[7] Uddin et al. (2017) utilized DL for FER, showcasing the effectiveness of unsupervised feature learning in FER. Their approach highlights the importance of leveraging unsupervised learning for feature extraction in FER.[8] Zhao et al. (2021) proposed Geometry-aware FER via Attentive Graph CNN, emphasizing the importance of incorporating geometric information for accurate expression recognition. Their method demonstrates the efficacy of graph CNN in FER tasks.[9] Zhang et al. (2023) introduced a Transformer-based Multimodal Emotional Perception model for dynamic FER, highlighting the importance of leveraging multimodal information. Their study emphasizes the significance of transformer-based architectures in FER.[10]. Wang et al. (2023) proposed Pose-Aware FER Assisted by Expression Descriptions, showcasing the importance of incorporating pose information for accurate expression recognition. Their approach demonstrates the effectiveness of pose-aware models in FER tasks.[11] Zhu et al. (2023) introduced Knowledge-conditioned Variational Learning for one-class FER, highlighting the importance of leveraging knowledge for robust expression recognition. Their study underscores the significance of incorporating prior knowledge into FER models.[12] Ma et al. (2023) proposed a Transformer-augmented Network with Online Label Correction for FER, emphasizing the importance of online learning for adaptive expression recognition. Their approach demonstrates the efficacy of online label correction in improving FER accuracy.[13] Kommineni et al. (2021) introduced a hybrid feature extraction technique for accurate FER, showcasing the importance of combining multiple features for robust expression recognition. Their study highlights the effectiveness of hybrid feature extraction in FER tasks.[14] Liu et al. (2023) proposed Uncertain FER via Multi-task Assisted Correction, emphasizing the importance of addressing uncertainty in FER. Their approach demonstrates the effectiveness of multi-task learning in handling uncertain expressions.[15] Karnati et al. (2023) introduced a Blended Feature Attention Network for FER in-the-wild, emphasizing the importance of attention mechanisms for handling complex facial expressions. Their study highlights the significance of attention-based models in real-world FER scenarios.[16] Li et al. (2023) proposed Cross-domain FER via Contrastive Warm Up and Complexity-aware Self-training, showcasing the importance of domain adaptation for robust expression recognition. Their approach demonstrates the efficacy of contrastive learning in cross-domain FER tasks.[17] Chen et al. (2023) introduced Cross-domain Sample Relationship Learning for FER, highlighting the importance of capturing sample relationships for domain adaptation. Their approach demonstrates the effectiveness of sample relationship learning in cross-domain FER scenarios.[18] Gao et al. (2024) proposed Adaptive Global-Local Representation Learning and Selection for Cross-Domain FER, emphasizing the importance of adaptive feature learning for cross-domain expression recognition. Their method demonstrates the efficacy of global-local representation learning in cross-domain FER tasks.[19]Vo et al. (2020) introduced Pyramid with Super Resolution for in-the-wild FER, showcasing the importance of super-resolution techniques for handling low-resolution facial images. Their approach demonstrates the effectiveness of super-resolution in real-world FER scenarios.[20]Dominguez-Catena et al. (2024) proposed Metrics for Dataset Demographic Bias, emphasizing the importance of evaluating dataset bias for fair FER. Their study underscores the significance of addressing demographic bias in FER datasets.[21] Qin et al. (2023) introduced Swinface, a multi-task transformer for FER, showcasing the importance of multi-task learning for comprehensive facial analysis. Their method demonstrates the efficacy of transformer-based architectures in multi-modal facial analysis tasks.[22] Zhao et al. (2023) proposed DFME, a new benchmark for dynamic facial micro-expression recognition, highlighting the importance of addressing temporal dynamics for accurate micro-expression recognition. Their benchmark provides a comprehensive evaluation framework for dynamic FER tasks.[23] Ye et al. (2024) introduced Dep-FER, a method for FER in depressed patients based on voluntary facial expression mimicry, showcasing the importance of personalized FER for mental health monitoring. Their study demonstrates the efficacy of personalized FER models in identifying emotional cues in depressed patients.[24]

3. MOTIVATION

The motivation behind this paper stems from the profound impact that FER technology can have on various aspects of our lives. In today's digital age, where human-computer interaction is becoming increasingly prevalent, FER holds immense potential to revolutionize numerous domains, ranging from healthcare and security to entertainment and communication. Understanding human emotions through

facial expressions is crucial for developing empathetic and intuitive human-computer interfaces. By accurately interpreting facial cues, machines can better comprehend user emotions, leading to more personalized and responsive interactions. This not only enhances user experience but also opens up new avenues for applications such as virtual assistants, emotion-aware educational tools, and therapeutic interventions. Furthermore, in fields like healthcare, FER can play a pivotal role in mental health monitoring and diagnosis. Detecting subtle changes in facial expressions can aid in the early detection of conditions like depression, anxiety, and autism spectrum disorders, facilitating timely interventions and improved patient outcomes. In the realm of security, FER technology can bolster surveillance systems by enabling the detection of suspicious or abnormal behavior based on facial cues. This can help in crime prevention, crowd management, and border security, enhancing public safety and security measures. Moreover, FER holds promise in entertainment and gaming industries, where immersive experiences can be created based on real-time analysis of user expressions. From interactive storytelling to emotion-driven gameplay mechanics, FER has the potential to elevate entertainment experiences to new heights. Overall, the motivation behind this paper lies in harnessing the transformative power of FER to address societal needs, improve human-computer interaction, and pave the way for innovative applications across various domains.

4. PROPOSED METHODOLOGY

4.1 DenseNet121 + GRU:

The model is designed to incorporate DenseNet121 as its core architecture. It includes a reshape layer to adapt the output for sequential processing, followed by the addition of two GRU layers, each with 256 and 128 units. These layers aim to capture temporal dependencies within facial data effectively. Furthermore, a dense layer with 256 units and ReLU activation facilitates feature aggregation. Finally, a dense layer with softmax activation, comprising seven units for emotion classification, completes the model. This approach enables feature extraction via DenseNet121 and sequential data processing through GRU layers, facilitating accurate emotion recognition from facial images. Proposed Architecture summary is in fig 1.

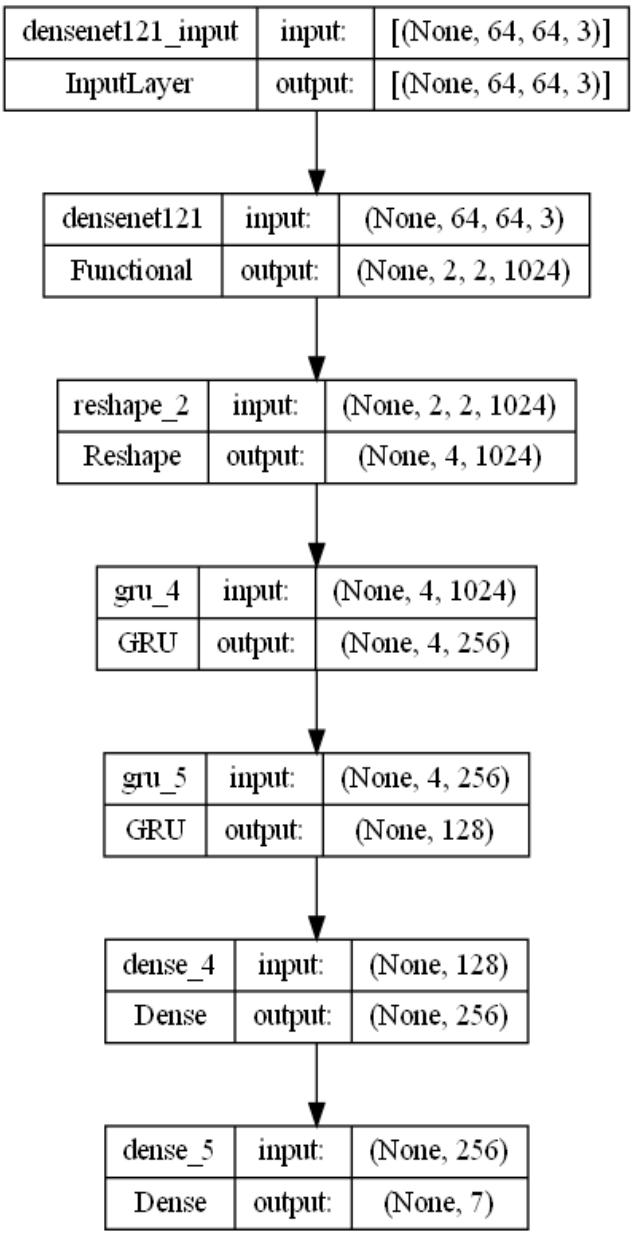

Fig 1. Architecture Summary using DenseNet121 + GRU

4.2 VGG16 + GRU:

The model integrates the VGG16 pre-trained model with a GRU (Gated Recurrent Unit) layer to analyze emotions depicted in facial images. Initially, VGG16 is utilized as a feature extractor after removing its top layers. Following this, a Global Average Pooling layer aggregates spatial information, while a Reshape layer introduces a temporal dimension for sequential processing. Then, a GRU layer captures temporal patterns inherent in the data. Finally, a Dense layer with softmax activation enables multi-class classification. This architecture effectively combines VGG16 for feature extraction and GRU for sequential analysis. Proposed Architecture summary is in fig 2.

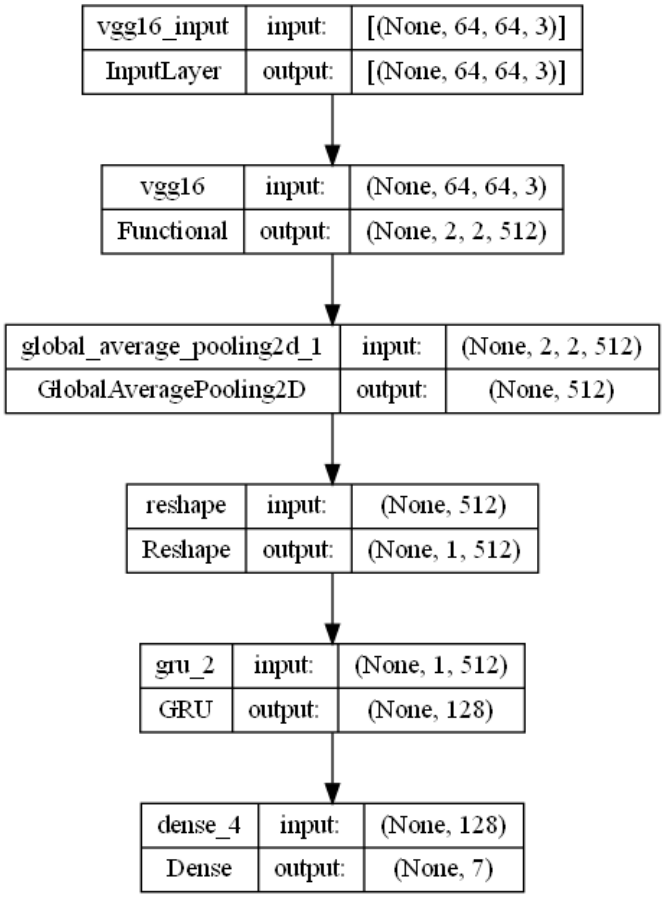

Fig 2. Architecture Summary using VGG16 + GRU

4.3 Xception + GRU :

The model architecture blends the Xception framework with a GRU layer to analyze facial images and forecast emotions. Initially, Xception is employed to extract features, succeeded by a MaxPooling2D layer to downsize spatial dimensions. Following this, the output is flattened and reshaped into a 3D tensor to accommodate the GRU layer, crucial for capturing temporal nuances. Further feature extraction is facilitated through a Dense layer employing ReLU activation, alongside dropout regularization to prevent overfitting. Ultimately, a Dense layer using softmax activation predicts emotional states. This amalgamation optimizes both feature extraction and temporal modeling, thereby refining the accuracy of emotion classification. Proposed Architecture summary is in fig 3.

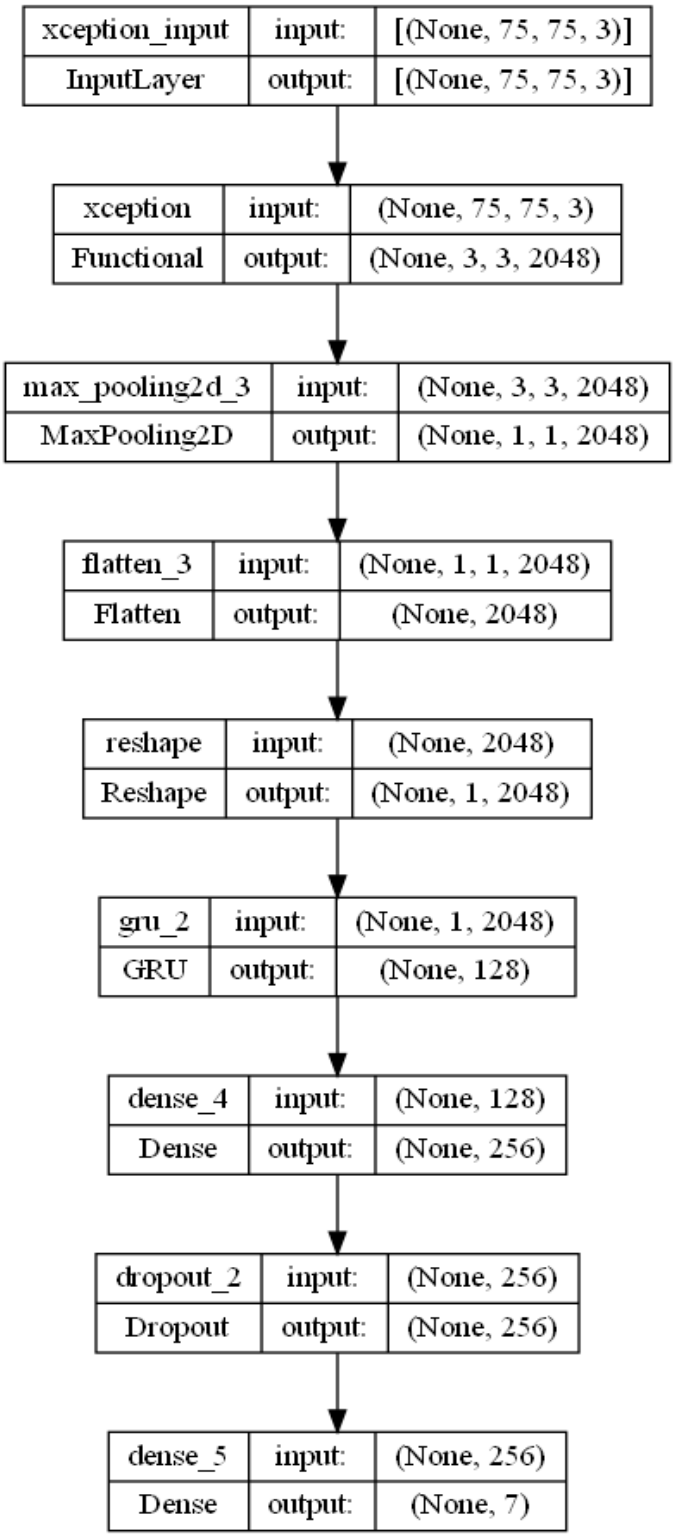

Fig 3. Architecture Summary using Xception + GRU

4.4 Dataset Description

The dataset employed in our study, namely the "OAHEGA: EMOTION RECOGNITION DATASET," encompasses RGB images portraying cropped facial expressions denoting six

distinct emotions: Happy, Angry, Sad, Neutral, Surprise, and Ahegao. In its raw form, the dataset initially comprised around 18,000 images. However, through meticulous data preprocessing and cleaning procedures, we refined and distilled the dataset to a more streamlined set of approximately 17,000 images. This curation process ensured a high-quality dataset, poised for robust training of facial emotion recognition models. The dataset's origin lies in a comprehensive amalgamation of sources, including social media platforms, frames extracted from YouTube videos, and integration of data from established repositories like IMDB and AffectNet. Structured for ease of use, the dataset is packaged as a zip file, featuring folders categorizing each emotion, complemented by a data.csv file providing image paths and corresponding emotion labels. This curated dataset stands as a valuable and diverse resource, poised to enhance the effectiveness and generalization capabilities of facial emotion recognition models. The reference for the dataset is [26].

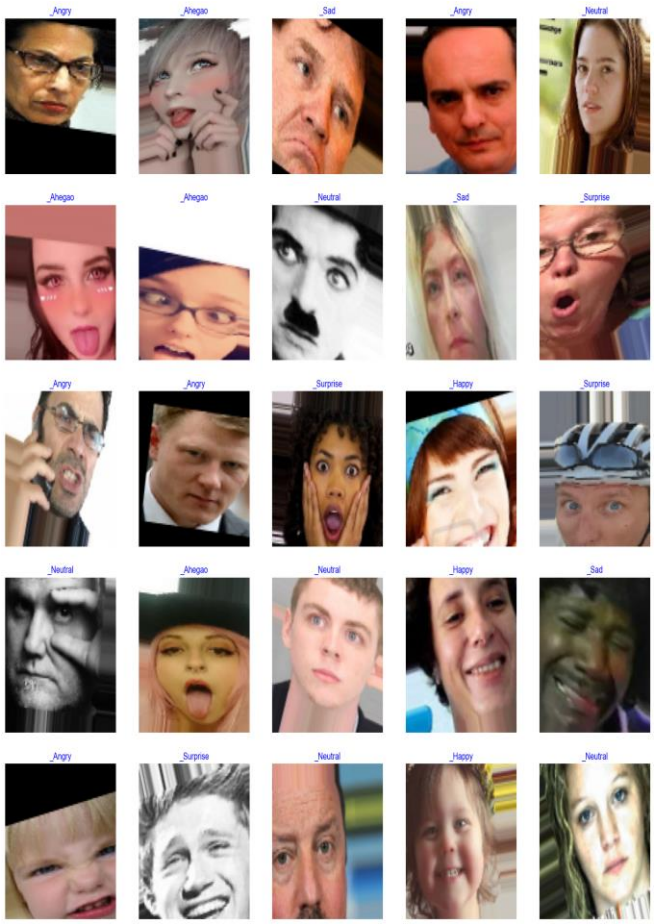

Fig 4. Sample Dataset Image

5. Results Analysis

Meticulous evaluation of emotion detection models across various classifiers, as depicted in the table below:

Table 1: Results based on Facial Emotion Recognition Image Dataset

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DenseNet121 + GRU	99.65	98.96	100	99.57
VGG16 + GRU	97.58	96.95	99	98
Xception + GRU	98.27	97.65	99.23	98.47

Table 1 provides a comprehensive examination of the performance metrics for different DL classifiers in the context of FER. DenseNet121 combined with GRU achieved exceptional results, demonstrating near-perfect accuracy (99.65%) and F1-Score (99.57%), with perfect recall (100%) indicating robust classification capabilities. DenseNet121 coupled with GRU demonstrates outstanding performance, achieving nearly flawless accuracy (99.65%) and F1-Score (99.57%), alongside perfect recall (100%), highlighting its robust classification capabilities. VGG16 integrated with GRU delivers commendable results, particularly in recall (99%), with an accuracy of 97.58%, precision of 96.95%, and F1-Score of 98%. Xception with GRU demonstrates strong performance across all metrics, achieving an accuracy of 98.27%, precision of 97.65%, recall of 99.23%, and F1-Score of 98.47%, indicating its capability in handling diverse datasets and complex patterns. Accuracy curve and roc curve for DenseNet121 + GRU model are shown in fig 5 & 6.

Fig 5. Accuracy curve using DenseNet121 + GRU

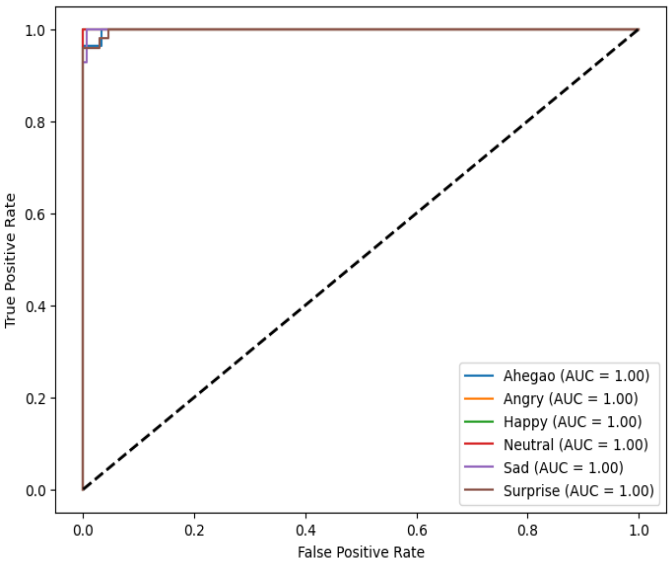

Fig 6. ROC Curve using DenseNet121 + GRU

6. CONCLUSIONS

This research has delved into the realm of FER, leveraging the power of DL to decipher the intricate nuances of human expressions. In conclusion, our exploration of FER using various DL models has unveiled insightful findings regarding their efficacy. Our analysis reveals that DenseNet121 coupled with GRU demonstrated exceptional performance, achieving impressive evaluation metrics on FER image dataset, showcasing its robust capability in accurately discerning facial emotions. Both VGG16 integrated with GRU and Xception with GRU delivered noteworthy performance across all evaluated metrics, demonstrating their versatility and effectiveness across varied datasets.

Moving forward, our study identifies several promising avenues for future research. Exploring ensemble methods that integrate the strengths of multiple models could potentially enhance overall performance in FER systems. Furthermore, investigating transfer learning techniques, wherein pre-trained models are adapted to specific FER datasets, may mitigate the need for extensive annotation efforts and improve generalization to unseen instances. Additionally, efforts to enhance the interpretability of DL models in FER tasks hold promise in providing deeper insights into decision-making processes critical for real-world applications. Moreover, considering the temporal dynamics of facial expressions and incorporating multimodal data, such as audio and text inputs, may further enhance the accuracy and robustness of emotion recognition systems. Lastly, exploring real-time implementations of these models on edge devices could facilitate deployment in diverse applications, including human-computer interaction, virtual reality, and affective computing. Overall, our study lays the foundation for future research endeavors aimed at advancing the field of FER and its practical applications.

REFERENCES

[1] Nan, Yahui, et al. "A-MobileNet: An approach of facial expression recognition." Alexandria Engineering Journal 61.6 (2022): 4435-4444.

[2] Wen, Zhengyao, et al. "Distract your attention: Multi-head cross attention network for facial expression recognition." Biomimetics 8.2 (2023): 199.

[3] Li, Bin, and Dimas Lima. "Facial expression recognition via ResNet-50." International Journal of Cognitive Computing in Engineering 2 (2021): 57-64.

[4] Pham, Luan, The Huynh Vu, and Tuan Anh Tran. "Facial expression recognition using residual masking network." 2020 25Th international conference on pattern recognition (ICPR). IEEE, 2021.

[5] Ma, Fuyan, Bin Sun, and Shutao Li. "Facial expression recognition with visual transformers and attentional selective fusion." IEEE Transactions on Affective Computing 14.2 (2021): 1236-1248.

[6] Wang, Kai, et al. "Region attention networks for pose and occlusion robust facial expression recognition." IEEE Transactions on Image Processing 29 (2020): 4057-4069.

[7] Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor. "Going deeper in facial expression recognition using deep neural networks." 2016 IEEE Winter conference on applications of computer vision (WACV). IEEE, 2016.

[8] Uddin, Md Zia, et al. "Facial expression recognition utilizing local direction-based robust features and deep belief network." IEEE Access 5 (2017): 4525-4536.

[9] Zhao, Rui, et al. "Geometry-aware facial expression recognition via attentive graph convolutional networks." IEEE Transactions on Affective Computing 14.2 (2021): 1159-1174.

[10] Zhang, Feifei, Mingliang Xu, and Changsheng Xu. "Weakly-supervised facial expression recognition in the wild with noisy data." IEEE Transactions on Multimedia 24 (2021): 1800-1814.

[11] Zhang, Xiaoqin, et al. "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild." IEEE Transactions on Circuits and Systems for Video Technology (2023).

[12] Wang, Shangfei, et al. "Pose-aware facial expression recognition assisted by expression descriptions." IEEE Transactions on Affective Computing 15.1 (2023): 241-253.

[13] Zhu, Junjie, et al. "Knowledge conditioned variational learning for one-class facial expression recognition." IEEE Transactions on Image Processing (2023).

[14] Ma, Fuyan, Bin Sun, and Shutao Li. "Transformer-augmented network with online label correction for facial expression recognition." IEEE Transactions on Affective Computing 15.2 (2023): 593-605.

[15] Kommineni, Jenni, et al. "Accurate computing of facial expression recognition using a hybrid feature extraction technique." The Journal of Supercomputing 77.5 (2021): 5019-5044.

[16] Liu, Yang, et al. "Uncertain facial expression recognition via multi-task assisted correction." IEEE Transactions on Multimedia (2023).

[17] Karnati, Mohan, et al. "Facial expression recognition in-the-wild using blended feature attention network." IEEE Transactions on Instrumentation and Measurement (2023).

[18] Li, Yingjian, et al. "Cross-domain facial expression recognition via contrastive warm up and complexity-aware self-training." IEEE Transactions on Image Processing (2023).

[19] Chen, Dongliang, et al. "Cross-Domain Sample Relationship Learning for Facial Expression Recognition." IEEE Transactions on Multimedia (2023).

[20] Gao, Yuefang, et al. "Adaptive Global-Local Representation Learning and Selection for Cross-Domain Facial Expression Recognition." IEEE Transactions on Multimedia (2024).

[21] Vo, Thanh-Hung, et al. "Pyramid with super resolution for in-the-wild facial expression recognition." IEEE Access 8 (2020): 131988-132001.

[22] Dominguez-Catena, Iris, Daniel Paternain, and Mikel Galar. "Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[23] Qin, Lixiong, et al. "SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation." IEEE Transactions on Circuits and Systems for Video Technology (2023).

[24] Zhao, Sirui, et al. "DFME: A New Benchmark for Dynamic Facial Micro-expression Recognition." IEEE Transactions on Affective Computing (2023).

[25] Ye, Jiayu, et al. "Dep-FER: Facial Expression Recognition in Depressed Patients Based on Voluntary Facial Expression Mimicry." IEEE Transactions on Affective Computing (2024).

[26] (https://www.kaggle.com/datasets/sujaykapadnis/emotion-recognition-dataset) (Kaggle Dataset)