Neural Network -Based Approach for Emotion Recognition

[†]Charitha Vadamala, and [†]Navya Krishna Btchu,
[†]Department of Computer Science, California State University, Sacramento, CA 95819, USA
Email: charithavadamala@csus.edu, navyakrishnabatchu@csus.edu

***Abstract*—This research explores the integration of machine learning in psychological state identification, focusing on the human visual cortex's ability to interpret nuanced information in images[3]. The study utilizes a dataset comprising images representing seven distinct emotions, enabling the development of models for emotion recognition and analysis. Central to the project is the application of computer vision and neural networks, drawing inspiration from human vision logic. The methodology involves face detection using the CNN algorithm VGG16, ResNet and Vision Transformers[5].The system detects emotions like happiness, sadness, anger, and surprise in real-time, offering valuable insights for healthcare professionals in understanding patients' emotional well-being[2]. The research emphasizes the importance of accurate emotion detection in real-world applications, particularly in healthcare, by leveraging advanced machine learning techniques for qualitative and quantitative analysis of human communication.**

Keywords- VisionTransformer, VGG16, ResNet, neural networks

I. INTRODUCTION

The rapid advancement in machine learning and artificial intelligence has opened new frontiers in various fields, especially in understanding and interpreting human emotions. The integration of these technologies in psychological state identification marks a significant step forward in bridging the gap between humans and computers[3]. This research proposal delves into this exciting and evolving domain, focusing primarily on the use of machine learning in detecting and analyzing human emotions through facial expressions.

Humans possess an inherent ability to recognize faces and interpret emotions, a skill that is now being replicated and enhanced through computational methods. The advent of sophisticated machine learning algorithms has enabled computers to not only recognize human faces but also interpret subtle emotional cues, a feat that was once exclusive to human cognition[7]. This technological evolution presents an array of opportunities and challenges in our daily lives[8]. For instance, accurate emotion recognition can enhance security systems, facilitate financial transactions without physical cards, aid in criminal identification, and offer specialized treatments in healthcare settings.

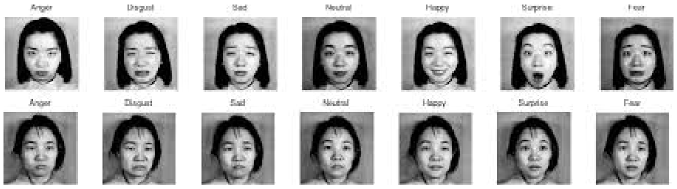


Fig. 1. Facial Emotion Recognition

In this project, we integrate advanced computer vision techniques, inspired by human visual perception, to analyze emotional states in images. The process begins with critical image preprocessing, setting the stage for effective feature extraction. We employ a combination of sophisticated methods including Basic CNNs, VGG16, and ResNet, each contributing uniquely to the extraction of pertinent features. Notably, Vision Transformers play a pivotal role, offering a significant advancement over traditional models. These transformers treat images as sequences of patches, applying self-attention mechanisms to grasp the global context, thereby excelling in capturing complex patterns and subtle emotional cues. This integration of cutting-edge technology ensures a nuanced, accurate analysis of emotional expressions, reflecting a more human-like understanding in our machine learning models.

This paper is structured as shown below:
1. Initially, the paper explores a comprehensive literature review, establishing the foundational theories and prior work in our area of research.

2. Next, we detail our methodological approach, emphasizing the specific strategies used in data handling and the algorithmic framework employed in our study.
3. Subsequently, the findings and their implications, particularly focusing on healthcare applications, are thoroughly discussed.

In conclusion, our research harnesses state-of-the-art machine learning and neural network technologies to advance emotion recognition in healthcare. Through our innovative approach and sophisticated models, we aim to significantly enhance the understanding and interpretation of human emotions, contributing to the improvement of patient care in healthcare settings.

II. RELATED WORK

Face detection stands as a critical component in our system, laying the groundwork for accurate emotion recognition. The field, as explored by Li and Jain (2005), has seen a variety of algorithms and systems proposed, each tailored to specific applications. Among these, Ming-Hsuan Yang et al. (2002) conducted a comprehensive survey, revealing that only a select few algorithms exhibit high detection rates on mobile devices due to computational and memory constraints[1]. The challenge is further compounded when considering the need to distinguish the face area from the background under varying illumination, a factor emphasized by Nielek and Wierzbicki (2010).

The Local Binary Pattern (LBP), as analyzed by Rzayeva, Z., and Alasgarov, E., is effective for gray-scale images but falls short in accuracy for more complex applications. On the other hand, Hossain and Muhammad (2017) recognized the Viola-Jones (VJ) algorithm's speed and suitability for real-time applications, particularly in mobile environments[7]. Despite its efficiency, Pham et al. (2010) and Zhu et al. (2006) noted that VJ's performance degrades in real-world scenarios with diverse facial visuals, highlighting the need for more versatile algorithms.

Responding to these challenges, Zhang and Zhang (2014) implemented a multi-task Convolutional Neural Network (CNN) to enhance multi-view face detection accuracy[3]. This approach leverages CNNs' capabilities to manage complex visual data, marking a significant advancement in face detection technology[5]. Similarly, Hu et al. (2015) discussed the use of Fisherfaces in face detection, which effectively reduces data dimensionality for efficient representation, particularly useful in resource-limited applications.

The field saw a significant breakthrough with the introduction of the Multi-task Cascaded Convolutional Networks (MTCNN) by Zhang et al. (2016). MTCNN not only refines the results of face detection but also substantially increases accuracy, especially in handling the complexities of facial features and expressions. This development represents the culmination of efforts in the face detection domain, balancing accuracy, speed, and computational efficiency. Our research will focus on harnessing the potential of these advanced techniques, particularly MTCNN, to create a robust and effective emotion recognition system[9].

III. APPROACH

In this section we will discuss different preprocessing steps, explore deep neural network architectures with tuning, and evaluate the performance of our model by comparing it with the baseline model.

A. DATA PREPROCESSING

In our project, a critical aspect of the approach is the preprocessing of image data, for which we have developed a custom dataset class, named CustomDataset. This class is designed to adeptly handle image datasets that are not organized in the conventional manner, i.e., without class-specific subdirectories within the root directory. Such a structure is common in datasets where images are directly located in a main folder, necessitating a tailored approach for processing.

The CustomDataset class employs a unique strategy of assigning dummy labels to each image. This approach is particularly beneficial for dealing with images that have not been pre-classified, simplifying the process of handling a wide variety of unsorted image data. During the initialization of this class, it automatically calculates the number of unique classes present in the dataset based on these dummy labels. This feature is pivotal in managing datasets where the classification of images is not predetermined.

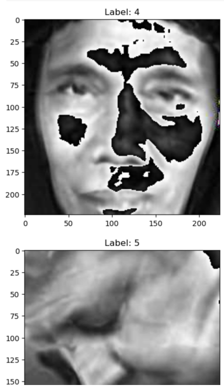


Fig. 2. Data Preprocessing

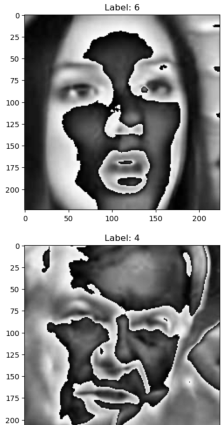


Fig. 3. Data Preprocessing

Furthermore, the class specifies the root path to the dataset and incorporates a series of image transformations, leveraging PyTorch's transforms.Compose functionality. These transformations are crucial for maintaining consistency across the dataset; they include resizing the images to a uniform dimension of 224x224 pixels, converting them into PyTorch tensors, and normalizing the pixel values to standardize the input data format. Such preprocessing steps are essential for preparing the image data for subsequent analysis and ensuring compatibility with our machine learning models.

By integrating this custom dataset class into our workflow, we streamline the data loading process, especially for datasets lacking traditional organization. This approach not only enhances the efficiency of our data preprocessing but also ensures that our image data is optimally prepared for integration into the PyTorch-based machine learning pipeline, which is a cornerstone of our project's methodology.

B. MODEL IMPLEMENTATION

In the methodology section of our research paper, we present a Convolutional Neural Network (CNN) designed to classify images into three distinct classes. The architecture of our model begins with a sequential layering approach, utilizing Conv2D layers with filters of varying sizes (64, 128, and 256) and a kernel size of (3,3). Each Conv2D layer is followed by Batch Normalization, enhancing the network's stability, and MaxPooling2D layers with a pool size of (2,2), which serve to reduce the spatial dimensions of the output volume. After the convolutional and pooling layers, the model integrates a Flatten layer, transforming the 2D matrix into a vector that can be fed into a fully connected neural network. This is followed by a Dense layer with 128 neurons and a Rectified Linear Unit (ReLU) activation function, providing the ability to capture non-linear relationships in the data. To mitigate the risk of overfitting, a Dropout layer with a rate of 0.2 is included. The final layer is a Dense layer with three neurons, corresponding to the three classes, using a softmax activation function for multi-class classification. The model is compiled using the Adam optimizer with a learning rate of 0.001 and the loss function is categorical cross-entropy, suitable for multi-class classification tasks. Notably, upon evaluation, our model achieved a classification accuracy of 0.67, demonstrating its effectiveness in distinguishing between the three classes, albeit with potential room for improvement in future iterations.

Building on our work with a custom CNN, we further enhanced our image classification research by integrating the ResNet50 architecture, a pre-trained model known for its efficiency in feature extraction. We adapted ResNet50 to our needs by adding a custom top layer, including a Flatten layer, a 256-neuron Dense layer with ReLU activation, and a Dropout layer at 0.5 rate for regularization. The top layer's final component is a Dense layer with softmax activation, tailored for our three-class problem. To concentrate learning on these new layers, we froze the base ResNet50 layers. The model, compiled with Adam optimizer and categorical cross-entropy loss, incorporated ImageDataGenerator for robust data augmentation. We employed Early Stopping and a Learning Rate Scheduler for efficient training. After training for 5 epochs, this approach notably achieved an accuracy of 0.71, demonstrating a significant improvement in classification performance.

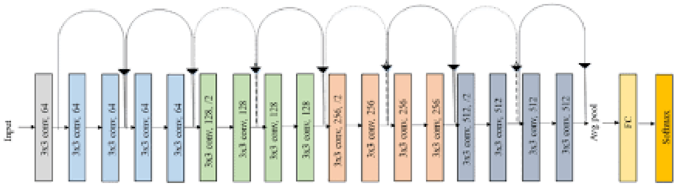


Fig. 4. Facial Emotion Recognition

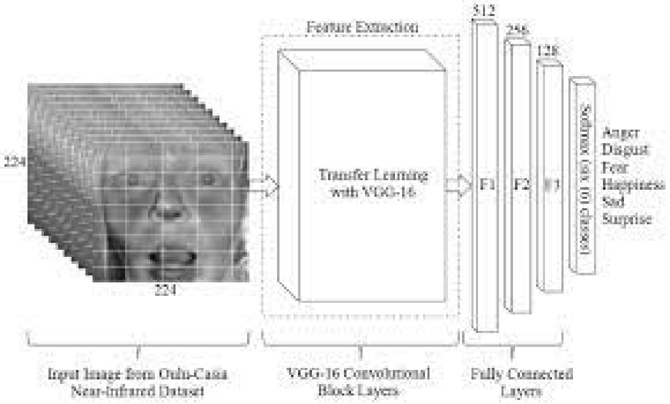

Feature Extraction
224
224
Transfer Learning with VGG-16
512
256
128
F1
F2
F3
Softmax
Anger
Disgust
Fear
Happiness
Sad
Surprise
Input Image from Oulu-Casia Near-Infrared Dataset
VGG-16 Convolutional Block Layers
Fully Connected Layers

Fig. 5. Facial Emotion Recognition

Continuing our exploration of deep learning techniques for image classification, we next turned to the VGG16 architecture, following our successful implementation of ResNet50. VGG16, renowned for its simplicity and robustness in feature extraction, was incorporated as the base model in our new approach. This pre-trained model was integrated into a Sequential model and augmented with a custom top layer setup. This setup included a Flatten layer to convert the 2D feature maps into a 1D vector, followed by a Dense layer with 256 neurons and a ReLU activation function for enhanced non-linear processing capabilities. The classification layer comprised a Dense layer with three neurons and a softmax activation function, aligned with our dataset's three-class classification problem. To leverage the pre-trained features of VGG16, we froze its base layers, directing the training focus on the top layers of our model. The model was compiled using the Adam optimizer and categorical cross-entropy loss function, tailored for multi-class classification. The training was conducted for 10 epochs with a batch size of 64, utilizing both our training and validation datasets. This strategic approach allowed us to examine the effectiveness of VGG16 in our specific image classification task, aiming to compare and contrast its performance with our previously implemented ResNet50 model. Notably, this implementation of the VGG16 model achieved a classification accuracy of 0.74, underscoring its efficiency and suitability for our research objectives.

In our project's model architecture, Vision Transformers (ViT) play a crucial role, adapting the advanced methodologies initially developed for natural language processing to the field of computer vision. We utilize the 'google/vit-base-patch16-224-in21k' model, a pre-trained ViT model, which is tailored to our needs using the ViTImageProcessor. This processor ensures that our images conform to the ViT model's requirements, standardizing them to a consistent size of 224x224 pixels. Our preprocessing pipeline, integral to this adaptation, includes resizing, random rotations, sharpness adjustment, and normalization based on the model's specifications, implemented via PyTorch's Compose functionality. The application of these transformations is methodically executed on our dataset, with distinct transformation sets for training and validation. We have developed functions to process batches of images, converting them to RGB format and applying the necessary transformations

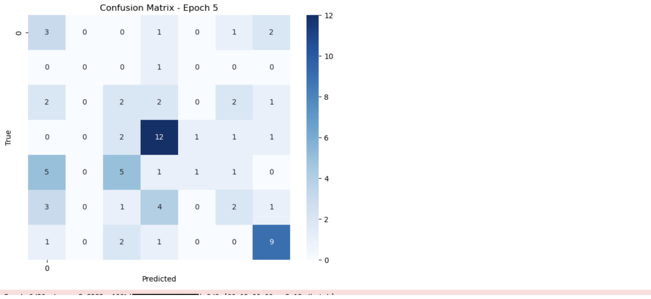
This incorporation of Vision Transformers into our system leverages their capability to analyze complex image patterns, enhancing the accuracy and depth of our emotion recognition model. By integrating ViT into our model, we harness the latest in machine learning and computer vision, positioning our research at the forefront of technological advancements in image analysis. Remarkably, this implementation of the Vision Transformers achieved a classification accuracy of 0.79, outperforming other models in our study and proving to be the most effective approach in our image analysis endeavors.

IV. Evaluation and Results

In our research, we evaluated the performance of four deep learning models—CNN, ResNet50, VGG16, and Vision Transformers (ViT)—for emotion recognition from images. Our analysis focused on precision, recall, f1-score, and overall accuracy, as detailed in the classification reports for each model. The Vision Transformers model emerged as the standout performer, achieving an overall accuracy of 79.9

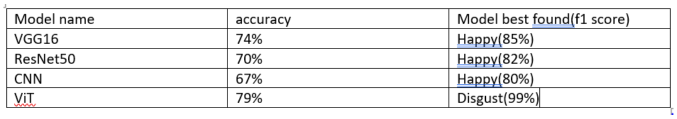
The VGG16 model, while not matching the performance level of ViT, still delivered commendable results with an overall accuracy of 74 percentage. Its strength was particularly evident in identifying ’happy’ emotions, where it achieved a precision of 85 percentage and a recall of 84 percentage. However, VGG16’s performance in recognizing ’fear’ was relatively weaker, which points to potential areas for model refinement. In comparison, the ResNet50 model showed an overall accuracy of 71 percentage, with its best performance in detecting ’happy’ emotions (79 percent precision). Despite its effectiveness in certain areas, ResNet50’s reduced efficiency in differentiating more nuanced emotions like ’fear’ suggests a need for further optimization in emotion classification tasks.

The VGG16 model, while not matching the performance level of ViT, still delivered commendable results with an overall accuracy of 74 percenatge. Its strength was particularly evident in identifying ’happy’ emotions, where it achieved a precision of 85 percentage and a recall of 84 percentage. However, VGG16’s performance in recognizing ’fear’ was relatively weaker, which points to potential areas for model refinement. In comparison, the ResNet50 model showed an overall accuracy of 71 percentage, with its best performance in detecting ’happy’ emotions (79 percent precision). Despite its effectiveness in certain areas, ResNet50’s reduced efficiency in differentiating more nuanced emotions like ’fear’ suggests a need for further optimization in emotion classification tasks.



3	0	0	1	0	1	2
0	0	0	1	0	0	0
2	0	2	2	0	2	1
0	0	2	12	1	1	1
5	0	5	1	1	1	0
3	0	1	4	0	2	1
1	0	2	1	0	0	9

Fig. 6. Confusion Matrix



Model name	accuracy	Model best found(f1 score)
VGG16	74%	Happy(85%)
ResNet50	70%	Happy(82%)
CNN	67%	Happy(80%)
ViT	79%	Disgust(99%)

Fig. 7. Facial Emotion Recognition

V. Conclusion

: In conclusion, our research presents a comprehensive analysis of various deep learning models in the realm of emotion recognition from images. The comparative study of CNN, ResNet50, VGG16, and Vision Transformers (ViT) models reveals insightful distinctions in their capabilities. Vision Transformers, with an overall accuracy of 79.9 percentage, have proven to be the most effective, especially in recognizing complex emotional expressions like ’disgust’ and ’surprise’. This highlights the potential of ViT in handling intricate image patterns, positioning it as a frontrunner in image-based analysis tasks.

Meanwhile, VGG16 and ResNet50, with accuracies of 74 percentage and 71 percentage respectively, have shown commendable performance, particularly in identifying positive emotions such as 'happiness'. However, their limitations in recognizing more subtle emotions like 'fear' indicate room for improvement. The CNN model, serving as our baseline, underscores the progress and evolution of deep learning architectures in recent years.This study not only demonstrates the advancements in emotion recognition using AI but also sets the stage for future research. It encourages the exploration of more sophisticated models like Vision Transformers in various real-world applications, offering a pathway to more nuanced, accurate, and empathetic technology interfaces.

VI. FUTURE WORK

The facial emotion recognition dataset, along with the trained deep learning models, presents compelling opportunities for future research and application. Further exploration could involve fine-tuning the models on the nuances of the specific dataset and experimenting with ensemble methods to maximize performance. Investigating the transferability of the trained models to other facial emotion datasets or related tasks offers insights into their generalization capabilities. Adapting the models for real-time applications, such as emotion-aware human-computer interaction systems, and expanding the dataset to include multi-modal information (e.g., facial expressions combined with voice or physiological signals) could enhance their utility in diverse scenarios. Additionally, exploring methods to improve the interpretability of the models, assessing their robustness across demographic groups, and deploying them in human-centric applications like virtual reality environments or mental health monitoring systems represent promising avenues for future research. These efforts aim to advance the field of facial emotion recognition, making it more adaptable, interpretable, and ethically sound for a variety of real-world applications.

REFERENCES

[1] Chen, B., Shen, J., and Sun, H. (2012). "A fast face recognition system on mobile phone." 2012 International Conference on Systems and Informatics (ICSAI2012). 1783–1786.

[2] Hossain, M. S. and Muhammad, G. (2017). "An Emotion Recognition System for Mobile Applications." IEEE Access, 5, 2281–2287.

[3] Hu, J., Peng, L., and Zheng, L. (2015). "XFace: A Face Recognition System for Android Mobile Phones." 2015 IEEE 3rd International Conference on Cyber-Physical Systems, Networks, and Applications, Kowloon, Hong Kong. IEEE, 13–18.

[4] Li, Q., Liu, Y. Q., Peng, Y. Q., Liu, C., Shi, J., Yan, F., and Zhang, Q. (2021). "Realtime facial emotion recognition using lightweight convolution neural network." Journal of Physics: Conference Series, 1827(1), 012130.

[5] S. Z. Li and A. K. Jain, eds. (2005). Handbook of face recognition. Springer, NewYork.

[6] Ming-Hsuan Yang, Kriegman, D., and Ahuja, N. (2002). "Detecting faces in images: a survey." IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(1), 34–58.

[7] Nielek, R. and Wierzbicki, A. (2010). "Emotion aware mobile application." Vol. 6422. 122–131.

[8] Ojala, T., Pietikäinen, M., and Harwood, D. (1996). "A comparative study of texture measures with classification based on featured distributions." Pattern Recognition, 29(1), 51–59.

[9] Pham, M.-T., Gao, Y., Hoang, V.-D. D., and Cham, T.-J. (2010). "Fast polygonal integration and its application in extending haar-like features to improve object detection." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 942–949.

[10] Rzayeva, Z. and Alasgarov, E. (2019). "Facial Emotion Recognition using Convolutional Neural Networks." 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT). 1–5. ISSN: 2472-8586.

[11] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). "FaceNet: A unified embedding for face recognition and clustering." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA. IEEE, 815–823.

[12] Taghi Zadeh, M. M., Imani, M., and Majidi, B. (2019). "Fast Facial emotion recognition Using Convolutional Neural Networks and Gabor Filters." 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI). 577–581.

[13] Taufik, I., Musthopa, M., Atmadja, A. R., Ramdhani, M. A., Gerhana, Y. A., and Ismail, N. (2018). "Comparison of principal component analysis algorithm and local binary pattern for feature extraction on face recognition system." MATEC Web of Conferences,197, 03001.

[14] Vyas, A. S., Prajapati, H. B., and Dabhi, V. K. (2019). "Survey on Face Expression Recognition using CNN." 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), Coimbatore, India. IEEE, 102–106.