# Face Detection and Recognition Using OpenCV and Vision Transformer

Mr. Krish Kumar
Department of Information Technology
National Institute of Technology,
Karnataka, India
krishraj1252@gmail.com

Mr. Nilesh Pingale
Department of Information Technology
National Institute of Technology,
Karnataka, India
nileshpinglenilu@gmail.com

Ms. Bhawana Rudra
Department of Information Technology
National Institute of Technology,
Karnataka, India
bhawanarudra@nitk.edu.in

***Abstract*— Face recognition technology is vital in the real world with diverse applications. It is primarily used for security, law enforcement, personalization, healthcare, and education. Face recognition systems use biometric features like facial landmarks, texture, and shape to identify and verify individuals. The suggested approach employs a transformer-based architecture that solely relies on self-attention and does not utilize Convolutional Layers. This design choice enables the model to be trained efficiently with minimal computational power and fewer parameters than a CNN. The application of Vision Transformer (ViT) in various computer vision tasks has been highly successful, making it a state-of-the-art approach. Given its superior performance, we are interested in exploring whether ViT can enhance the accuracy of sheep face recognition.**

**In this paper, we show that ViT can be a useful technique for facial recognition. Since there was no predefined dataset for face recognition, a PCI dataset was built for this investigation. Along with the PCI dataset, two more well-known datasets, AT&T and 5_Celebrity, we used to examine performance. In our model was seen that ViT could identify human faces on the PCI dataset with a 99% accuracy rate and perform much better than other face recognition algorithms like Eigenface, FisherFace, and LBPH.**

***Keywords*— *OpenCV, EigenFace, FisherFace, Transformer, LBHP Algorithm, Vision Transformer.***

## I. INTRODUCTION

Face identification and detection are crucial computer vision tasks with many applications in various fields. Achieving high accuracy in a variety of situations, such as those with varied lighting conditions, head positions, facial expressions, and occlusion, is one of the most difficult parts of face identification and recognition. In recent years, a variety of algorithms and methods, including conventional statistical-based approaches, machine learning-based strategies, and deep learning-based strategies, have been presented to solve these difficulties.

Using Linear Discriminant Analysis (LDA), Fisherface, a variation of Eigenface, enhances the separation between various classes of faces in the feature space. This results in better performance for facial recognition tasks with significant lighting, pose, and expression variations.

Vision Transformer (ViT) the proposed model is a more recent approach to computer vision that uses self-attention mechanisms to learn meaningful representations of visual input. ViT has attained cutting-edge performance in several computer vision tasks, including object identification and picture categorization.

In terms of performance, Vision Transformer generally outperforms the other approaches, but it also requires significantly more computational resources. LBPH, Eigenface, and Fisherface are all relatively simple and efficient methods that can be used for real-time facial recognition tasks. However, they may not perform as well as ViT in more challenging scenarios with larger datasets or more complex visual input.

Delbiaggio et.al. [11] said Computer vision problems have been tackled using various neural network architectures, including convolutional neural networks (CNN) and recurrent neural networks (RNN) with encoder-decoder structures. While CNN-based architectures like VGG and Resnet have been successful in addressing complicated image recognition issues, the Sequential data necessitates the utilization of multiple layers to effectively capture long-term dependencies and may suffer from information loss when using max-pooling. RNNs, on the other hand, can learn long-term memory dependencies but can't be trained concurrently because of their linear structure.

In natural language processing (NLP), The Transformer, an attention-based design, has gained widespread use and has proven to be highly effective in tasks such as translation and sentiment analysis. As a result of its success, researchers in computer vision have expressed interest in employing similar self-attention models. Google has utilized the Vision Transformer (ViT) in computer vision tasks, leading to advancements in image recognition prototypes such as ImageNet. [15]

Vision Transformer (ViT) is a neural network design that processes information using the self-attention mechanism to develop image data. ViT has emerged as a promising alternative to achieve comparable or even better results with fewer parameters Ramachandran et. al. [18].

In summary, ViT offers a different approach to image recognition that detects long-range dependencies that can be captured via self-attention methods in image data. Alexander Kolesnikov et.al. [21]. While CNNs are still the most widely used architecture for image recognition, ViT has shown promising results and may become a more popular alternative in the future.

The following is the sequence in which the paper has been set up: Introduction (I), Existing Work (II), Methodology, Proposed Approach (IV), Result and Discussion (V), and Conclusion (VI).

## II. EXISTING WORK

The impact of image detection and recognition in today's world is immense, with numerous ongoing studies focused on developing and improving algorithms for image recognition.

G. S. Nagpal et.al. [1] In computer vision, the HaarCascade classifier and the LBPH are two frequently employed methods for identifying and detecting faces. A combination strategy that uses the Haar Cascade classifier for face detection in an image and LBPH for facial identification has been suggested in certain research papers. The fusion of these two methods results in a comprehensive facial recognition system that can proficiently detect faces within an image and subsequently identify them based on their distinctive characteristics.

Face recognition and identification were performed using the LBPH algorithm and OpenCV. According to Kumuda et al. [24], OpenCV, an open-source computer vision software package, is a well-known framework for image and face recognition. The LBPH algorithm is a well-liked technique for texture-based face detection.

A. M. Jagtap et.al. [2] have examined the efficacy of various techniques, including LBPH, Eigenface, Fisherface, and Haar-like features. These algorithms are widely used and have been applied directly to recognize faces. These aim to determine which algorithm is the most effective for face recognition.

Bussa et.al. [3] Upgrading the attendance system in colleges, schools, and organizations. OpenCV has been developed, incorporating a camera for capturing input images, an algorithm for detecting and encoding faces, and a system for marking attendance and converting it into a PDF file. The system was trained using the faces of authorized students, and the cropped photographs that result are then kept as a database with the appropriate labels. The LBPH algorithm is applied to the database to extract features.

X. Zhao et.al. [4] Despite being a straightforward method for face recognition (LBPH) algorithm occasionally struggles to sustain recognition rates when confronted with issues including lighting variations, expression changes, and orientation deviations. A modified LBPH algorithm, known as the Median Local Binary Pattern Histogram (MLBPH), we proposed to overcome these limitations. In the MLBPH algorithm, a pixel's gray value is substituted with the median value of the surrounding sampling value.

Due to blurriness, poor lighting, limited resolution, and irregular lighting, identify and recognize humans using video surveillance cameras. A. Ahmed et.al. [5] proposed a remedy for these issues entails the development of the LR500 database, which has 500 photos of each person. To eliminate noise and align the positioning of each image, the normalization approach was performed on all photos.

In the Suma et.al. [6] paper, The LBPH and Viola-Jones algorithms have been utilized to improve the effectiveness of face recognition. The Viola-Jones algorithm was chosen for its high precision and real-time processing capability. Using the LBPH, facial features associated with the detected faces are retrieved from the live stream. The Euclidean distance classifier was used for face identification. The experimental results suggest an accuracy range of 85% to 95% after implementation and testing.

To effectively recognize faces, having the appropriate hardware in place with the existing LBPH algorithm is essential. P. Bhatia et.al. [7] proposed, using a Raspberry Pi III microcontroller, The Raspbian Stretch operating system will be used, and the door will have a camera for facial recognition and a stepper motor to unlock it. Each person would be photographed between 100 and 150 times from various perspectives and lighting setups for the image database. The LBPH model will be used for face recognition.

Ahsan et.al. [8] Face recognition in difficult circumstances, such as dim lighting, variable illumination, and bad weather, is still difficult and needs more study. Eigenface, Fisherface, and LBPH algorithms have been evaluated in prior research for facial recognition in such settings. The results indicated that LBPH was the most dependable across diverse lighting conditions, making it the most efficient algorithm for the task. The optimal settings for the four LBPH parameters— radius, neighbors, grid, and threshold value—in terms of accuracy and computing time have not yet been determined through experimentation. Additionally, LBPH's performance in challenging circumstances needs to be further acknowledged. The Lamar University database (DARTC) and the 5_celebrity dataset are used in this paper's in-depth experiment to assess the four LBPH parameters. A new approach called BMC-LBPH (Bilateral Median Convolution-Local Binary Pattern Histogram) has been introduced and tested in wet conditions using a UAV that has four vision sensors. The experiment's findings show that BMC-LBPH works better than traditional LBPH methods, with accuracy rates 5_celebrity dataset (65%,), L.U. dataset (98%,), and rainy weather (8%), and all of these results were achieved in real-time.

## III. DATASET COLLECTION

There is no existing dataset available, especially for the testing of face recognition applications. The dataset is generated in place and stored in a database using SQL Lite.
As a result, a fresh facial dataset called "DARTC" was generated for this research. To evaluate performance, the DARTC dataset was combined with two other commonly used datasets, namely “AT&T and 5_Celebrity” which are shown in TABLE I. The 5_Celebrity dataset was acquired from Kaggle, a well-known website for machine learning resources.

TABLE I. DATASET USED IN PAPER

| Dataset Name | Participants | Images | Size |
|---|---|---|---|
| **AT&T** | 40 | 400 | 92 × 112 |
| **5-Celebrity** | 5 | 120 | Not Equal |
| **DARTC** | 20 | 300 | 500 × 500 |

## IV. METHODOLOGY

The effectiveness of Face Detection and Recognition relies on the ability of the model to extract image features accurately. Although models such as LBPH have shown promising results, they sometimes fall short due to insufficient image feature extraction. As a solution, we suggest the use of a Vision Transformer.

We proposed an advanced and latest model for extracting image features, which passed to the next model phase to detect the face. ViT is similar to the Transformer but has a

different model architecture that works on the image. That's why it is also called an alternative to CNN.

**There are overall three stages of model implementation**:

1. **Making dataset at runtime:** whenever the program captures the face will write that in the folder. Before grabbing the face, tell the scripts whose face it is. We need an identifier (ask the user for their UserId). Then it will capture 100 snapshots of each image and write them in the folder.
2. **Train the Recognizer:** In this step, the ViT model will take the images produced in the previous phase and extract the features from them.
3. **Detection Algorithm:** It will detect the user by fetching the data from the database.

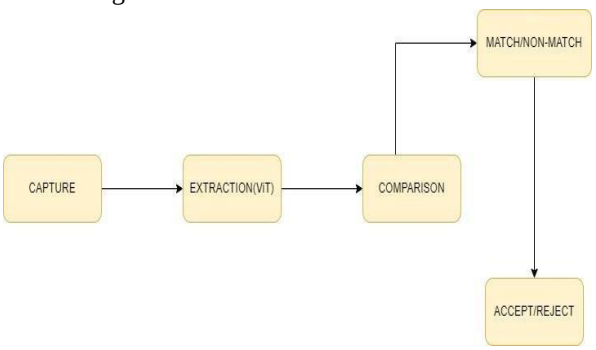


Fig. 1. Stage of identification

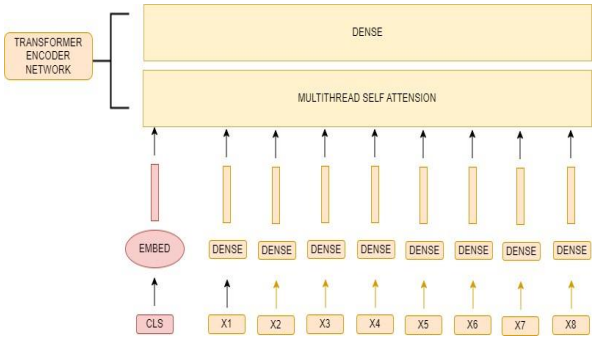


Fig.2 Encoder phase of Vision Transformer model

*A. Mathematical Intuition Behind Vision Transformer*

It is necessary to first restructure the input picture $Z \in T^{M. \times Y \times D}$. into flattened image patches before using the conventional Transformer for two-dimensional (2D) image processing. $Z_r \in T^{P \times (R \times S \times E)}$, where (J, Y) represents the real picture size, D is the number of channels, (R, R) is the image patches' x- and y-coordinates, and P = JYR-2 gives the total number of patches.

They are processed in a fixed dimension using all layers after being linearly projected into E-dimensional vectors. To produce the transformer encoder's input sequence, the merged categorization tokens and patches are subjected to positional embedding. The encoder block has both multi-layer perceptrons and multi-head self-attention layers Scaled dot-product attention, commonly referred to as self-attention, uses queries set to S =.ZYS, a set of keys M = $ZY_M$, and values set X.= $ZY_X$ are calculated, where $Z \in T^{(p+m)\times F}$ represents the input sequence, and $Y_S$., $Y_T$, and $Y_X$. are trainable weights with a vector size of e. The computation of self-attention is then performed.:

$$\text{Attention}(S, M, X) = \text{softmax}\{SM^V/\sqrt{e}\} \quad (1)$$

The softmax function was applied to each row of the matrix. I head are appended to self-attention in MSA in the following ways::

$$\text{MSA}(S, M, X) = \text{join}(\text{head}_0, \dots, \text{head}_{i-1})Y \quad (2)$$

$$\text{head}_{i-1} = \text{Attention}(S_{i-1}, M_{i-1}, X_{i-1}) \quad (3)$$

The size of the sequence provided by each head is used to create i sequences which are then rearranged into a (0+1) × fj single sequence. The MLP (multi-layer perceptron) then adjusts the orientation into (0+1) × F. The following describes how the transformer encoder processed the image.

$$A=[U_{\text{class}}; \text{MLP}(U^1_r); \dots ; \text{MLP}(U^0_r)+G_{\text{pos}} \quad (4)$$

$$A'_n=\text{MSA}(\text{Norm}(A_{m-1})+ Z_{m-1}) \quad (5)$$

$$A_n = \text{MLP}(\text{Norm}(A_n')) + A'_n \quad (6)$$

$$H = \text{Norm}(A^0_n) \quad (7)$$

The mathematical foundation shown above is all behind the Vision Transformer involves the model's ability to process images by dividing them into patches, flattening them, performing vectorization, and then sending the resulting input sequence tokens to the encoder. There are some stages to the walkthrough for face recognition shown in Fig. 1. And the encoder phase of Vision Transformer is also shown in Fig, 2.

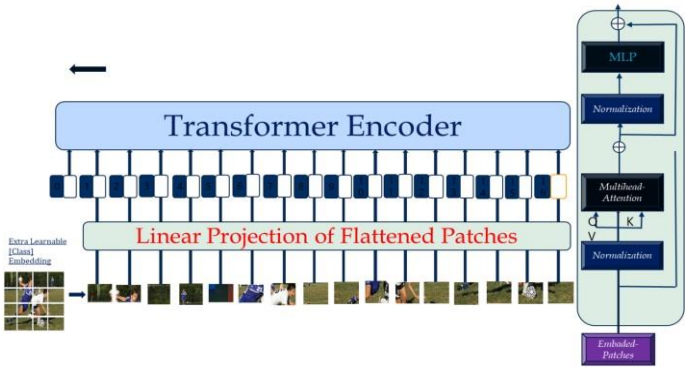


Fig. 3. Complete Vision Transformer Architecture

## V. RESULT AND DISCUSSION

The results of the investigation are presented in the section that follows. The outcomes show that the suggested method

outperformed other deep learning models in terms of accuracy.

We utilized a PC with an Intel(R) Core(T.M) i5-8265U CPU operating at 1.60GHz and 8 Gb of RAM to implement the proposed framework. Four situations were used in the accuracy analysis, and they are shown below.

### *A. **Eigenfaces***

The Eigen Face algorithm is being used for the first time to detect faces in this situation. To identify facial pictures, üge Çarıkçı et.al. [10] used eigenfaces and principal component analysis (PCA). PCA has several advantages, including low noise sensitivity and dimensionality reduction capabilities. The Euclidean Distance is the distance between the eigenvectors of the eigenfaces. While a high distance suggests that the model needs more training to identify the topic, a short distance shows that the topic has been discovered.

### *B. **Fisher Face***

The Fisher Face method was employed in the second instance to find the face. To identify the typical middle face for each class, the Fisher Face (F.F.) technique modifies the Eigenface (E.F.) method. Please consult Delbiaggio's paper (2017) for a thorough understanding of the Eigenface approach Delbiaggio et.al. [11].

### *C. **Local Binary Pattern Histogram***

LBPH operator, which was first proposed as a texture descriptor, labels every pixel value in the image. To do this, a 3 3 neighborhood surrounding a center pixel value that is later processed as a binary integer is thresholded. Sánchez López et.al. [12] (Lopez, 2010) provides more information on the subject.

### *D. **Vision Transformer (proposed model):***

This is the last scenario where we discussed the proposed model(ViT).

The Vision Transformer model generally follows these steps:

a. Data Preparation: The first step is to prepare the dataset for training the model, which includes tasks such as data cleaning, resizing, normalization, and augmentation.

b. Tokenization: The next step is to convert the image data into tokens, which are sequences of numbers that the transformer model can process.

c. Patch Embedding: The image tokens are then grouped into patches of fixed size, and each patch is transformed into a feature vector using a learned embedding function.

d. Positional Encoding: To incorporate the spatial relationship between the patches, positional encoding is added to the feature vectors.

e. Multi-Head Self-Attention: The feature vectors are fed into the multi-head self-attention mechanism, which enables the model to simultaneously attend to different parts of the image.

## VI. COMPARISON AMONG EXISTING MODEL AND PROPOSED MODEL

The four cases listed above are EigenFace, FisherFace, LBPH, and VIT(proposed). The experiment comprised comparing the performance of the three different approaches Eigenface (E.F.), Fisherface (F.F.), LBPH, and Vision Transformer (ViT) on three different datasets ( AT&T, 5_Celebrity, and DARTC ). To preserve an equal data ratio and prevent any potential data imbalance-related complications, 100 images were randomly chosen from each dataset, even though AT&T and DARTC each contain 400 and 250 images, respectively. The evaluation of the full outcome was conducted using k-fold cross-validation, where k = 10, and the results of nearly all folds were averaged to provide the conclusions. The statistical results were reported using the 95% confidence interval due to the small datasets Ahsan et.al. [25-26]. The following formulae were used to determine the execution length, F1 score, recall, accuracy, precision, and overall effectiveness:

- **Cross Validation:** One way to evaluate the model's performance is to pass a subset of the test data as input to the model.

  Popular methods:
  - Stratified k-fold
  - Leave one out
  - Validation set approach
  - Leave-p-out
  - k-fold

- **Accuracy:** To check model accuracy, It is a ratio of correct predictions to all possible predictions.

- **Precision**: To validate accurate True Prediction.

- **Recall**: Identify all relevant cases in a dataset.

- **F1-Score**: Quantifies the accuracy of a model by taking into account both its precision and recall scores.

TABLE II. COMPARISON TABLE

| MODEL | DATASET | ACCURACY | PRECISION | RECALL | F1-SCORE |
|---|---|---|---|---|---|
| EIGENFACE | AT&T | 95.677% | 0.9324 | 0.9634 | 0.967 |
| | 5_Celebrity | 56% | 0.523 | 0.5159 | 0.598 |
| | DARTC | 86.2% | 0.8513 | 0.8327 | 0.863 |
| FISHERFACE | AT&T | 96% | 0.9754 | 0.9633 | 0.973 |
| | 5_Celebrity | 61% | 0.6441 | 0.6132 | 0.619 |
| | DARTC | 84.53% | 0.8312 | 0.8112 | 0.858 |
| LBPH | AT&T | 97% | 0.986 | 0.9632 | 0.9879 |
| | 5_Celebrity | 72% | 0.737 | 0.7218 | 0.731 |
| | DARTC | 95% | 0.94 | 0.93 | 0.96 |
| **VIT(proposed)*** | AT&T | 98% | 0.982 | 0.985 | 0.989 |
| | 5_Celebrity | 97% | 0.977 | 0.9657 | 0.982 |
| | DARTC | 99% | 0.998 | 0.9865 | 0.997 |

## VII. CONCLUSION AND FUTURE WORK

This research objective is to find the effectiveness of the standard ViT for human face recognition tasks. The proposed framework utilizes an entirely self-attention-based transformer architecture, which doesn't rely on convolutional layers. This approach enables the model to be trained efficiently with minimal computational power and fewer parameters than CNN. Moreover, it leverages data augmentation to enhance the precision of the model. The outcomes demonstrated that integrating the ViT model with OpenCV and transformer structures, it provides a lightweight approach to achieving face recognition. Additionally, this model can be deployed on the Jetson Nano edge device to accomplish real-time and accurate recognition outcomes, hence propelling the progress of digital technology in agriculture and promoting the practical deployment of deep learning in precision agriculture.

***Future Work:*** Vision transformers have demonstrated significant potential in various computer vision tasks, including face recognition. To further enhance the performance of vision transformers in face recognition, several areas of research can be explored. Firstly, data augmentation methods can be investigated to enhance the performance of vision transformers on smaller datasets, as they heavily rely on a large amount of data. Secondly, optimizing the model architecture, such as reducing the number of parameters, can help address the high computation and memory requirements of vision transformers.

Transfer learning can also be utilized to fine-tune pre-trained vision transformers for specific face recognition tasks, which can significantly reduce the data required for training and improve the overall performance of the model. Additionally, research can be done on enhancing the robustness of vision transformers to adversarial attacks, which can negatively impact their performance.

Finally, since vision transformers require a substantial amount of time for training and inference, research can be conducted to improve the speed of vision transformers for real-time face recognition applications.

Overall, the future scope for vision transformers in face recognition is vast, and further research in these areas can lead to significant advancements in the field.

## REFERENCES


[1] G. S. Nagpal, G. Singh, J. Singh, and N. Yadav, "Facial Detection and Recognition using OpenCV on Raspberry Pi Zero," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018, pp. 945-950, doi: 10.1109/ICACCCN.2018.8748389.

[2] A. M. Jagtap, V. Kangale, K. Unune and P. Gosavi, "A Study of LBPH, Eigenface, Fisherface and Haar-like features for Face recognition using OpenCV," 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2019, pp. 219-224, doi: 10.1109/ISS1.2019.8907965..

[3] Bussa, Sudhir, et al. "Smart attendance system using OPENCV based on facial recognition." Int. J. Eng. Res. Technol 9.3 (2020): 54-59.

[4] X. Zhao and C. Wei, "A real-time face recognition system based on the improved LBPH algorithm," 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), Singapore, 2017, pp. 72-76, doi: 10.1109/SIPROCESS.2017.8124508. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] A. Ahmed, J. Guo, F. Ali, F. Deeba and A. Ahmed, "LBPH based improved face recognition at low resolution," 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2018, pp. 144-147, doi: 10.1109/ICAIBD.2018.8396183.

[6] Suma, S. L., and Sarika Raga. "Real time face recognition of human faces by using LBPH and Viola Jones algorithm." International Journal of Scientific Research in Computer Science and Engineering 6.5 (2018): 610.

[7] P. Bhatia, S. Rajput, S. Pathak and S. Prasad, "IOT based facial recognition system for home security using LBPH algorithm," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2018, pp. 191-193, doi: 10.1109/ICICT43934.2018.9034420.

[8] Ahsan, M.M.; Li, Y.; Zhang, J.; Ahad, M.T.; Yazdan, M.M.S. Face Recognition in an Unconstrained and Real-Time Environment Using Novel BMC-LBPH Methods Incorporates with DJI Vision Sensor. J. Sens. Actuator Netw. 2020, 9, 54. https://doi.org/10.3390/jsan9040054.

[9] Ahsan, M.M. Real Time Face Recognition in Unconstrained Environment; Lamar University-Beaumont: Beaumont, TX, USA, 2018.

[10] üge Çarıkçı, M., and Figen Özen. "A face recognition system based on eigenfaces method." Procedia Technology 1 (2012): 118-123.

[11] Delbiaggio, N. A Comparison of Facial Recognition’s Algorithms. 2017. Available online: https://www.theseus.fi/handle/1002 4/132808 (accessed on 20 April 2021).

[12] Sánchez López, L. Local Binary Patterns Applied to Face Detection and Recognition. 2010. Available online: https://upcommons.

[13] Kumar, Krish, et al. "The Hybrid Vision Transformer Approach for Hyperpigmentation Nail Disease Detection." Proceedings of Second International Conference on Sustainable Expert Systems: ICSES 2021. Singapore: Springer Nature Singapore, 2022.

[14] Sun, Chen, et al. "Revisiting unreasonable effectiveness of data in deep learning era." Proceedings of the IEEE international conference on computer vision. 2017.

[15] Rezvantalab, Amirreza, Samir Mitha, and April Khademi. "Alzheimer's Disease Classification using Vision Transformers." (2021).

[16] Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic routing between capsules." arXiv preprint arXiv:1710.09829 (2017).

[17] Wang, Xiaolong, et al. "Non-local neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[18] Ramachandran, Prajit, et al. "Stand-alone self-attention in vision models." arXiv preprint arXiv:1906.05909 (2019).

[19] Mahajan, Dhruv, et al. "Exploring the limits of weakly supervised pretraining." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[20] Xie, Qizhe, et al. "Self-training with noisy student improves imagenet classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[21] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In ECCV, 2020.Enrollment in local colleges, 2005

[22] Thomsen, Kenneth, et al. "Deep learning for diagnostic binary classification of multiple-lesion skin diseases." Frontiers in medicine 7 (2020): 604.

[23] Saranya, V., and A. Ranichitra. "Image Segmentation Techniques To Detect Nail Abnormalities." Scholar 2 (2017): 1.

[24] Kumuda, S. "An Image Pre-processing Method for Fingernail Segmentation." 2017 IEEE 2nd International Conference on Signal and ImageProcessing.

[25] Ahsan, M.M.; Ahad, M.T.; Soma, F.A.; Paul, S.; Chowdhury, A.; Luna, S.A.; Yazdan, M.M.S.; Rahman, A.; Siddique, Z.; Huebner, P.Detecting SARS-CoV-2 from Chest X-ray using Artificial Intelligence. IEEE Access 2021, 9, 35501–35513

[26] Ahsan, M.M.; Gupta, K.D.; Islam, M.M.; Sen, S.; Rahman, M.; Shakhawat Hossain, M. COVID-19 Symptoms Detection Based on NasNetMobile with Explainable A.I. Using Various Imaging Modalities.