Proceedings of the Fifth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2021)
IEEE Xplore Part Number: CFP21J88-ART; ISBN: 978-1-6654-3524-6

Detecting Pneumonia using Vision Transformer and comparing with other techniques

1st Khushal Tyagi
Computer Science and Engineering
College of Engineering Roorkee
Roorkee, Uttarakhand, India
khushal.tyagi.kt@gmail.com

2nd Gaurav Pathak
Computer Science and Engineering
College of Engineering Roorkee
Roorkee, Uttarakhand, India
gauravpathak95688@gmail.com

3rd Rahul Nijhawan
Computer Science and Engineering
University of Petroleum and Energy Studies
Dehradun, Uttarakhand, India
rahulnijhawan2010@gmail.com

4th Ankush Mittal
Computer Science and Engineering
Roorkee, Uttarakhand, India
dr.ankush.mittal.@gmail.com

2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-3524-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICECA52323.2021.9676146

Abstract*—Pneumonia is life-threatening. It's critical for infants, young children, elders, and people with health problems or enfeebles immune systems. However, someone who has been infected with coronavirus can get intense Pneumonia in each lung. The best way to stumble on Pneumonia is via chest X-ray. Radiotherapist is required for an examination of chest X-Ray. An automated pneumonia detection device would be helpful for early detection in far-off places. The proposed method makes it possible to train ViT models with enhanced performance. Nowadays, ViT is an alternative method of CNN in the field of computer vision. In this research, three models have been proposed, namely convolutional neural network (CNN), VGG16, and Visual Transformer were constructed. Statistical results are obtained after the comparison of all three models. Results indicate that ViT can identify Pneumonia with an accuracy of 96.45%. And also can be used to recognize other lung-related diseases. All the models were trained and tested on a dataset that contains standard chest X-Rays and pneumonia chest X-Rays.

***Index terms*—Pneumonia detection, Convolutional Neural Networks**

I. INTRODUCTION

More than 98% of children under the age of 5 years are dead due to Pneumonia in developing countries [1]. Pneumonia is the single most significant infectious cause of death in children worldwide. According to WHO, about 800,000 kids under the age of 5 years were killed by this disease in 2017 [2]. Pneumonia is an infectious disease in which infection causes the alveoli in one's lung to be filled with some fluid or pus, resulting in painful breathing and decreasing oxygen intake. And to detect Pneumonia, a precise examination of chest X-ray images is required by the radiographer or radiotherapist. That's why pneumonia detection is a time-consuming process, and a minor mistake can have an unbearable pay-off.

Though the most effective way of diagnosing Pneumonia is using chest X-Ray images, examining the X-ray for determining the location and extent of septic is very challenging as the appearance of Pneumonia in images of chest X-Ray can be very difficult blurry, which may give misleading results.

Computer Vision techniques are the most precise ways for chest X-Ray image examination to detect Pneumonia. CNN's have ruled in computer vision tasks so far. An image is based on the idea that one pixel is dependent on its neighboring pixels, and the next pixel is dependent on its immediate adjacent pixels (be it color, brightness, contrast, and so on). Different researchers developed many algorithms to recognize Pneumonia using different approaches like "ChexNet" [3], a CNN of 121 layers. Also, some more approaches like single-shot detectors and squeeze-and-extinction deep CNN [4]. Some researchers tried to combine and utilize some pretrained CNN models like AlexNet, VGG-19, etc.

From many studies, it is found that deep learning techniques are getting used to accomplish desired results on different sets of medical data like R. Nijhawan et al. [5] proposed a framework that utilizes a hybrid of CNN to extract features of images of other nails to detect different kinds of nail diseases. Also, D. Chandra et al. [6] proposed an architecture based on VGG-16 feature extraction for the detection of Progeria Syndrome in new-born babies.

In this paper, a Vision Transformer-based approach is proposed for examining chest X-Ray images for Pneumonia detection and compared it with CNN and VGG16 approaches on the same data. It was observed that ViT had outperformed the CNN and VGG16 approaches. To our knowledge, there is no other work till now that sought the use of Vision Transformer in the field of chest X-Ray image examination.

The work has been discussed in the following six sections: Introduction {no. 1} Literature Review {no. 2} Data-set {no. 3} Methodology and approaches {no. 4} Result and Discussion {no. 5} Conclusion {no. 6}

II. LITERATURE REVIEW

Many researchers have tried different computer vision techniques to detect Pneumonia using X-Ray images of human chests, e.g., Pranav Rajpurkar et al. [3] developed a set of rules named "ChexNet" that can stumble on Pneumonia from chest X-rays at a stage exceeding training radiologists. CheXNet is a 121-layer convolutional neural network skilled on ChestX-ray14, presently the largest publicly available chest X-ray dataset, containing over 100,000 frontal-view X-ray pictures with 14 sicknesses. Another research work by Tatiana Gabruseva et al. [4] advanced the computational method for pneumonia areas detection based totally on single-shot detectors, squeeze-and-extinction deep convolutional neural networks, augmentations, and multi-challenge getting to know. The proposed technique became evaluated inside the context of the Radiological Society of North America Pneumonia Detection venture, reaching one of the excellent effects within the venture. Dimpy Varshni et al. [7] appraised the functionality of pre-educated CNN models applied as function-extractors followed by one-of-a-kind classifiers for the unusual and everyday chest X-Rays analytically decided the most advantageous CNN version for the motive. The result of the proposed model is with an accuracy of 80.02%. Chouhan V. et al. [8] proposed an ensemble model that mixes outputs from all pretrained fashions, which outperformed person fashions, achieving the overall performance in pneumonia reputation. Their ensemble model reached an accuracy of 96.4% with a bear in mind today's 99.62.% on unseen facts from the Guangzhou girls and youngsters' clinical center data set. Hojjat Salehinejad et al. [9] solved the paucity problem of medical data by mixing the original chest X-Ray images with GAN generated chest X-Ray images then they applied DCNN on the dataset which improved the performance of classification at a great extent. Xiaosong Wang et al. [10] proposed a text-image embedding network to extract features and then they presented an auto-annotation framework which attained an appreciable accuracy of 0.9.

M. Toğaçar et al. [11] employed CNN as a function extractor and applied some of the present convolutional neural network models, such as AlexNet, VGG-sixteen, and VGG-19 to comprehend this particular project. They have got reduced blended functions with the feature selection method (mRMR). Suganya G et al. [12] combined VGG16 with Bi-directional LSTM to extract features of chest X-Ray images and performed classification using a fully connected layer for the diagnosis of tuberculosis and got appreciable accuracy of 97.76%. Qing Guan et al. [13] applied the VGG-16 DCNN model on cytological images to differentiate between papillary thyroid carcinoma and benign thyroid nodules and got a reasonable accuracy of 95% in patients. Defang Zhao et al. [14] proposed artificial data generation using forward and backward GAN and fed into multi-scale VGG16 model for feature extraction then they evaluated the accuracy on Lung Image Database Consortium and Image Database Resource Initiative dataset to be 95.24%.

In the medical field, Satish [27] in their research proposed a modified graph cut technique solving both accuracy and speed problems in conservative graph cut functions in diagnosing CT scan images for lung cancer disease.

III. DATASET

In this study, the Dataset used to train, test, and validate the models was 5,856 X-Ray images. 1,583 images are of Chest X-Ray of regular patients who don't have Pneumonia, and 4,273 images are of Chest X-Ray of patients having Pneumonia. All chest x-rays images were obtained from daily routine checkups of patients. For any further details of image quality, can refer [15]. Chest X-Ray images can be blurry; hence it can be difficult to detect pneumonia for the human eye. Simple phlegm can be misunderstood with pneumonic pus as the X-Ray image become hazy because of both reasons.

The test and train are divided into two categories: Pneumonia and routine chest x-rays. The model was trained and tested with data-set divided with 90% of the pneumonic chest X-Ray images and 84.7% of the normal/routine chest X-Ray images in the training set and 10% of the pneumonic chest X-Ray images and 15.3% of normal/routine chest X-Ray images in the testing set.

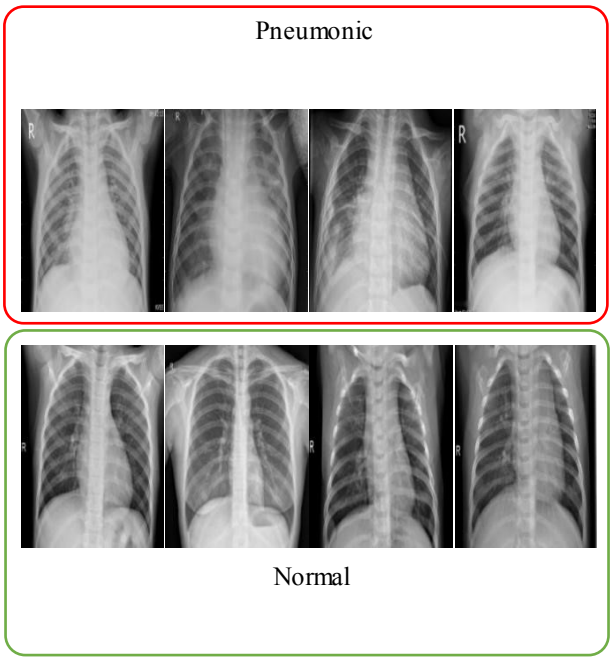

Fig 1. Showing Two Categories of images in Data-set

IV. METHODOLOGY

In this design, the similar approaches of the original Transformer is followed with some differences, which is trained on the ImageNet-21K Data set [16]. The input image is divided into 25 patches of 100×100, which linearly embeds

every patch of the image, adds position embed-dings, and fetches the sequences to an encoder.

Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz is used. Basic CNN, VGG-16, and Vision Transformer results are compared to find the best approach to detect Pneumonia.

A. VISION TRANSFORMER APPROACH

Nowadays, in Natural Language Processing (NLP) tasks, transformers have become a handy way. In computer vision, Vision transformer (ViT) implements a pure transformer model without convolutional blocks [17]. For many years CNN is used in image recognition. However, CNN has some drawbacks, A CNN is significantly slower due to an operation such as max pool, and ConvNet requires a large Dataset to process and train the neural network [18].

The model is proposed based on the Vision Transformer (Vit) approach to classify Pneumonia using a dataset of chest X-rays. Recently, Vision transformer [17] was preferred over CNN for large-scale computer vision datasets. Transformer architecture with self-attention allows ViT to integrate information across the entire image.

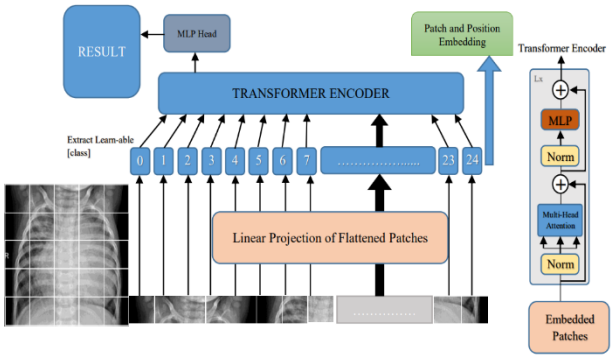

Fig 2. Vision Transformer Architecture

The image is broken into equal-sized patches. The small patches are also known as tokens. The series of permits is reshaped by 2D flattening into a vector format. Then a position embedding is added to the patch embedding to preserve positional information. The transformer encoder [18] consists of multi-head attention. The encoder contains self-attention layers. Embedded patches are connected to layer normalization in multi-head, and then again, layer normalization is connected to multi-layer perceptron blocks.

All the X-Ray images were resized to 250×250 pixels, then each image is broken down into 25 patches of 50×50 pixels each. These patches were then flattened and vectorized to feed into the transformer encoder network which adds positional encoding to the image vectors. A total of 6 transformer blocks are used with 8 number of heads in multi-head attention layer. Adam optimizer has been used. Parameters that are passed with their values in ViT model are given below:

- image_size – 250 – size of image in pixels.
- patch_size – 50 – size of each patch in pixels.
- channels – 3 – number of channels in image.
- num_classes – 2 – number of classes to classify.
- dim – 64 – last dimension of output tensor
- depths – 6 – total no. of transformation blocks
- heads – 8 – total no. of heads in multi-head attention layer
- mlp_dim – 128 – dimension of mlp layer

B. Convolutional Neural Networks approach

The Convolutional Neural Networks approach consists of multiple hidden layers which extract the information from an image. ReLU (Rectified Linear Unit) activation layer has been used. ReLU only passes values 0 for negative pixels. It introduces non-linearity to the network. Various filters are used in the pooling layer to identify different parts of the images. Then flattening is used to create a linear vector. The flattened matrix is fetched as input to the fully connected layers, used to classify the image [7, 21, 22, 23, 26].

C. VGG-16 approach

The data are pre-processed by re-sizing all images to 224×224 pixels after that, rescaling the pixel values by 1/255. Then, horizontal flip is applied to half of the pictures selecting randomly, followed by random shear transformations and zooming. Softmax function is used as activation function in output layer to predict a multinomial probability distribution.

The Sequential method is used as a sequential model has been created. A sequential model means that all the layers of the model will be arranged in sequence. Here, a VGG-16 pre-trained model trained on the "Imagenet" Dataset is used [16]. Then all the layers of the model are frozen to train. ADAM optimizer and learning rate decay are used to optimize the learning process.

V. RESULT AND DISCUSSION

The results of three different approaches are observed and evaluated. Therefore, the best result is obtained by comparing them.

Experiment No. 1

Convolutional neural networks is used in which four max pool layers, 1 Soft-max and 2 Rectified linear units (Relu) were

applied for better computational time to make it better to classify by non-linearity. In this model, an accuracy of 90.52% is achieved [23, 24, 25].

Experiment No. 2

VGG 16 CNN architecture is used. The images have been re-scaled by dividing the pixel values by 255. To maintain the uniform size of the image, the images are configured to shape (224, 224). ResNet50 model is used here as a base model for transfer learning. This model proposes an accuracy of 93.30%.

Experiment No. 3

Vision Transformer (ViT) is used to extract the features using attention layers and the model is trained in two classifications of datasets where the image is broken into 25 patches and then sequenced as linear embedding. The accuracy is 96.45% by using this technique.

After analyzing the results of all three models, it is found that ViT is better than CNN models. The primary trouble with CNNs, they fail to encode the spatial features. CNN does not consider the position of detecting characteristics concerning each other. In the Vision transformer, self-attention is used where it divides the image into small patches which are trainable and give importance to each part of the image and fetch into the Transformer alongside their positions. ViT implements a natural transformer model without the need for convolutional blocks. ViT is also more effective at doing complex tasks. Due to self-attention, transformer architecture can compute in a parallel manner to minimize computing time [19]. It can concurrently extract all the records needed from the input and its inter-relation, compared to CNN's. The result can be seen in the table. 1 of all three approaches and compared accuracies in the graph shown in fig. 3.

	ViT	CNN	VGG-16
Accuracy	0.9645	0.9052	0.933
Val Accuracy	0.8638	0.7885	0.8597
Loss	0.1092	0.1981	0.2633
Val Loss	0.1825	0.3799	0.7122

Table 1. Showing results of all three approaches

Table 1 shows that ViT gave an acceptable accuracy of 0.9645 on training data and accuracy of 0.8638 on unseen validation data with a fairly minimized value of cost function/loss. In the case of VGG-16, accuracy on train and validation data is acceptable at 0.933 and 0.8597, respectively, but the value of cost function/loss is considerably high and highly unacceptable. In the CNN approach, the accuracy and validation accuracy are also acceptable. The value of cost function/loss is comparably higher than ViT and lower than VGG-16. Cross Entropy Loss function has been used in all three approaches to calculate the loss.

A. ADVANTAGES OF VIT OVER CNN

ViT divides image into fixed size patches whereas CNN uses pixel arrays. In ViT patches are embedded according to their respective positions which leads to better results in feature extraction. Also ViT surpasses CNN in computational efficiency and accuracy.

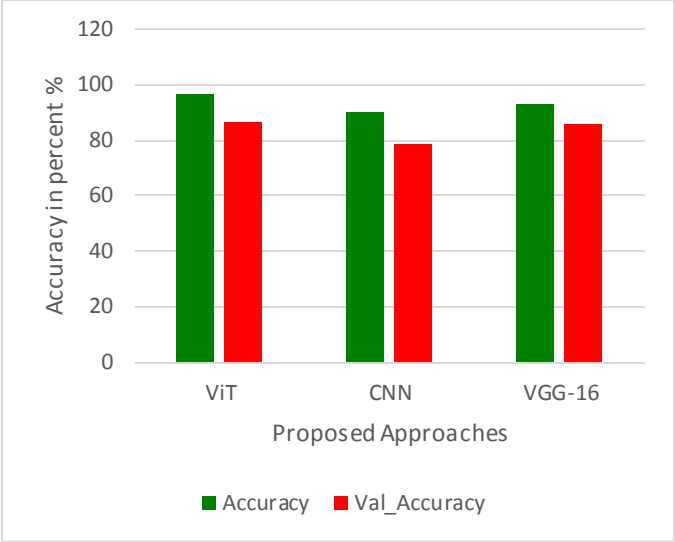

Fig 3. Graph showing comparison between accuracies of three approaches

B. LIMITATIONS

CNN's depends on the size of their filters and the number of convolutional layers used. Increasing the value of these hyper-parameters increases the complexity of the model, which can produce vanishing gradients or even models impossible to train. Residual connections and dilated convolutions have also been used to improve the receptive fields of these models, but the way convolutions operate over texts always presents limitations and trade-offs on the receptive field that it can capture.

Unlike CNN, ViT works on self-attention does not contain a convolutional layer. The performance of ViTs saturates fast when scaled to be more profound. More specifically, it is empirically observed that the attention collapse issue causes such scaling difficulty: as the Transformer goes deeper, the attention maps gradually become similar and even much the same after specific layers [19].

C. FUTURE WORK

In the field of chest x-ray diagnosis, not much work has been done in the Vision transformer. In the future, it can be beneficial for the detection of other diseases such as Pleural thickening, Covid-19, Edema, Effusion, Emphysema or Cystic Fibrosis, and even Cancer.

VI. CONCLUSION

In this paper, a Vision Transformer model is proposed for the early detection of Pneumonia to reduce the time-consuming chest X-ray evaluation process in far-off places. It can be seen that this approach of Vision Transformer gives comparable accuracy of 96.45% on the chest X-Ray data. Specialized radiology is the most crucial point for adequate diagnosis of any chest sac disease. It can prevent unfortunate outcomes in such far-off places.

REFERENCES

[1] Theodoratou E, Zhang JSF, Kolcic I, Davis AM, Bhopal S, et al. (2011) Estimating Pneumonia Deaths of Post-Neonatal Children in Countries of Low or No Death Certification in 2008. PLoS ONE 6(9): e25095. doi: 10.1371/journal.pone.0025095

[2] WHO URL: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[3] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning"; Cornell University arXiv:1711.05225, 14 Nov 2017.

[4] Tatiana Gabruseva, Dmytro Poplavskiy, Alexandr Kalinin "Deep Learning for Automatic Pneumonia Detection"; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 350-351.

[5] R. Nijhawan, R. Verma, Ayushi, S. Bhushan, R. Dua and A. Mittal, "An Integrated Deep Learning Framework Approach for Nail Disease Identification," 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2017, pp. 197-202, doi: 10.1109/SITIS.2017.42.

[6] D. Chandra, S. S. Rawat and R. Nijhawan, "A Machine Learning Based Approach for Progeria Syndrome Detection," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 74-78, doi: 10.1109/ISCON47742.2019.9036229.

[7] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan and A. Mittal, "Pneumonia Detection Using CNN based Feature Extraction," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-7, doi: 10.1109/ICECCT.2019.8869364.

[8] Chouhan, V., Singh, S.K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R. and De Albuquerque, V.H.C., 2020. A novel transfer learning based approach for pneumonia detection in chest X-ray images. Applied Sciences, 10(2), p.559

[9] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak and J. Barfett, "Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 990-994, doi: 10.1109/ICASSP.2018.8461430.

[10] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Ronald M. Summers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9049-9058.

[11] Toğaçar, M., et al. "A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models." Irbm 41.4 (2020): 212-222.

[12] Gutta, Jignesh Chowdary, G. Suganya, M. Premalatha, and K. Karunamurthy. "Class dependency based learning using Bi-LSTM coupled with the transfer learning of VGG16 for the diagnosis of Tuberculosis from chest x-rays." medRxiv (2021).

[13] Guan, Qing et al. "Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study." Journal of Cancer vol. 10,20 4876-4882. 27 Aug. 2019, doi:10.7150/jca.28769

[14] Zhao, Defang, Dandan Zhu, Jianwei Lu, Ye Luo, and Guokai Zhang. "Synthetic medical images using F&BGAN for improved lung nodules classification by multi-scale VGG16." Symmetry 10, no. 10 (2018): 519.

[15] Daniel Kermany; Kang Zhang; Michael Goldbaum, (2018) "Labled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, v2 published 06-01-2018.

[16] Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2009.5206848

[17] Alexey Dosovitskiy∗, Lucas Beyer∗, Alexander Kolesnikov∗, Dirk Weissenborn∗, Xiaohua Zhai∗, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby∗, ∗ equal technical contribution, † equal advising Google Research.

[18] Vaswani, Ashish, et al. "Attention is all you need." arXiv preprint arXiv:1706.03762 (2017).

[19] Zhou, Daquan, et al. "Deepvit: Towards deeper vision transformer." arXiv preprint arXiv:2103.11886 (2021).

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

[21] R. Nijhawan, H. Sharma, H. Sahni and A. Batra, "A Deep Learning Hybrid CNN Framework Approach for Vegetation Cover Mapping Using Deep Features," 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2017, pp. 192-196, doi: 10.1109/SITIS.2017.41.

[22] Nijhawan, R., Das, J., & Raman, B. (2018). A hybrid of deep learning and hand-crafted features-based approach for snow cover mapping. International Journal of Remote Sensing, 1–15. doi:10.1080/01431161.2018.1519277

[23] Nijhawan, R., Joshi, D., Narang, N., Mittal, A., & Mittal, A. (2018). A Futuristic Deep Learning Framework Approach for Land Use-Land Cover Classification Using Remote Sensing Imagery. Advances in Intelligent Systems and Computing, 87–96. doi:10.1007/978-981-13-0680-8_9

[24] S. Gupta, A. Panwar, S. Goel, A. Mittal, R. Nijhawan and A. K. Singh, "Classification of Lesions in Retinal Fundus Images for Diabetic Retinopathy Using Transfer Learning," 2019 International Conference on Information Technology (ICIT), 2019, pp. 342-347, doi: 10.1109/ICIT48102.2019.00067.

[25] Y. K. Arora, A. Tandon and R. Nijhawan, "Hybrid Computational Intelligence Technique: Eczema Detection," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 2472-2474, doi: 10.1109/TENCON.2019.8929578.

[26] S. S. Rawat, K. S. Rawat, V. Rawat and R. Nijhawan, "Neural Networks based Hand-crafted genetic learning approach to simulate Space Mario Game," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 1-5, doi: 10.1109/ICOSEC49089.2020.9215233.

[27] Sathish, Prof. "Adaptive Shape based Interactive Approach to Segmentation for Nodule in Lung CT Scans." Journal of Soft Computing Paradigm 2, no. 4: 216-225.