Vision Transformer Framework Approach For Melanoma Skin Disease Identification

Vikas Kumar Roy
Ruprecht Karl University of Heidelberg
Vasu Thakur
Ruprecht Karl University of Heidelberg
Nupur Goyal (drnupurgoyal10jun@yahoo.com)
Graphic Era Deemed to be University

Research Article

Keywords: Vision transformer (ViT), Machine Learning, Convolutional neural networks (CNNs), Skin cancer

Posted Date: February 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2536632/v1

Abstract

In the past few decades, skin diseases have been a hazardous issue because of more sophisticated and high-cost treatments. Identifying skin disease is still a challenging task for dermatologists. In reference to severe diseases like Melanoma, therapy in the initial stages is very important and effective to avoid skin cancer. This paper proposes an effective approach by using Vision Transformers (ViT) to detect Melanoma, which gives the accuracy of 99% on the test images. Authors considered the dataset, which is publicly available on Kaggle that comprises 1000 images and did the comprehensive study to get better results using ViT. The obtained results are compared with other state-of-the-art algorithms (VGG-19 and Inception-V3) to analyze the distinction between the proposed approach and other Convolutional Neural Network (CNN) Models.

1. Introduction

Skin is the exterior surface of the human body. There is a very high probability of getting infected with a polluted environment because every time the body comes in its contact. UV radiation is another source of skin cancer. People avoid skin-related diseases due to lack of medical knowledge. Skin disease like melanoma are hazardous cancer and most common in 75% of people and very fatal specifically if not diagnosed in early stages (Vijayalakshmi, 2019). Furthermore, cancer cells under our skin can't be diagnosed in the early stages unless these are mutated and infect the other body parts. At this specific stage, disease diagnosation is very complex. Melanoma is severe, it infects the internal body parts, and it becomes very difficult for dermatologists to diagnose it well. If it is treated at the early stages, the disease can be cured easily. The fundamental cause of this disease is the existence of Melanocytes. Dermoscopy is a conservative procedure to examine the composition of the skin to identify malignant conditions such as Melanoma. The 100x lens dipped in oil and incident light is used to observe the substructure of the skin. The precision of Dermoscopy is dependent on the Dermatologists.

In recent years, skin disease has become a research-oriented topic. Many researchers are focused on identifying skin cancer to detect Melanoma using various machine learning and artificial intelligence approach. In the past few years, Convolutional Neural Networks (CNNs) have been used to overcome the limitations of computer vision. Because of the linear functions of recurrent models, they cannot assist parallelization, which affects the long-term learning of the model. ResNet resolves the sophisticated identification of images on the ImageNet (Rawat et al., 2019). Although, CNNs architecture has limitations when it comes to dependencies in the sequential information. Moreover, loss of data can be caused by max-pooling (Deng et al., 2009). CNN based techniques have two major drawbacks- models are more focused on identifying non-global data rather than long-range correlation among images and CNN avoids local data by max-pooling and relative positional information. These limitations influenced the researchers to associate the self-attention mechanism with CNNs. In which they faced difficulty due to unavailability of the skin image dataset. The structure of the skin of an individual is relative.

Deep learning is a subset and advanced method of machine learning, which is essentially a neural network with three or more layer that work like a human brain which led to excessive focus on disease detection. (Cıcero et al., 2016) proposed a deep learning model to identify the melanoma disease and used ResNet as a feature extractor and detects skin lesion border through trained CNN, which results in an accuracy of 93.6%. They used to identify boundaries (Garnavi, 2011). Few procedures comprise upgraded color passages (Garnavi et al., 2010). The researchers made an effort to mark a similar problem with image recognition methods (Mengistu et al., 2015). (Celebi et al., 2009) proposed a non-local recognition procedure to detect melanoma regions. They performed the dermoscopic techniques through computer-aided diagnoses of Melanoma to evaluate and analyze the skin lesion. The main objective of the paper was automatic segmentation. (Celebi et al., 2007) proposed a methodological way to identify the skin lesions on dermoscopic images by separating the lesion from the background skin. The instability of the classifier is evaluated using Monte Carlo cross-validation. In (Celebi et al., 2007), the author used 564 images, and the result accuracy is 93.3%.

(Stolz, 1994) proposed an ABCD rule of dermatoscopy, which came out very reliable and allowed less computation. (Ng et al., 2005) proposed a methodology by using symmetric distances that intensify distinguishing potential methods. (Dreiseitl et al., 2001) evaluated the logistic regression technique, Artificial Neural Network (ANN), decision trees, and Support vector machines (SVM) to recognize the skin lesions as Melanoma. (Dreiseitl et al., 2001) represents the magnificent results through ROC curve analysis. (Gilmore et al., 2010) analyzed the melanoma disease using an SVM algorithm on 199 dermoscopic images and accomplished an AUC of 0.76. (Gilmore et al., 2010) proposed to integrate the SVM with the dermatological methodologies. (Menzies et al., 2005) described an automated tool to identify Melanoma at early stages. (Menzies et al., 2005) trained the instrument with a set of 2430 images comprising 382 melanoma images. (Manousaki et al., 2006) designed an image tool to distinguish Melanoma by developing a mathematical model to calculate the probability of Melanoma. Similarly, we've decided to observe and analyze the model using our proposed approach and represent a comparative analysis of ViT with other state-of-the-art algorithms (VGG-19 and Inception V3).

In this work, authors proposed a transformer structure with a self-attention mechanism called Vision Transformer (ViT) that uses a corps of Convolutional Neural Nets. In the past few years, it (Chhabra et al., 2020) has opted for large datasets, and applied a self-attention mechanism to the image patches to predict accurately. Several statistical measures have been computed to prove the efficacy of our proposed approach. The statical analysis and comparison of different state-of-art model like AUC, precision and accuracy have been done and achieved the results with best accuracy. The comparative analysis of the skin disease with deep learning techniques includes VGG-19, Inception-V3, and Vision Transformer (ViT) are provided.

This paper has arranged in the following order: Introduction (Section 1), Dataset (Section 2), Methodology (Section 3), Results and discussion (Section 4), Advantages of the proposed approach are described in Section 5 and conclusion are given in Section 6.

2. Dataset

We have acquired a dataset that is publicly available on Kaggle. The dataset consists of 1000 normal and malignant melanoma images. We've only considered the images of good quality to improve the accuracy of our proposed approach. Few samples of normal and diseased images are shown in Fig. 1.

The dataset is divided into two halves: training data comprises 80% images and 20% images for testing data. The number of images available in our dataset is displayed in the following Table 1.

Table 1
Number of images in Dataset

Skin	Number of Images
Melanoma	500
Normal	500

3. Methodology

Our methodology is comprised in three phases: Acquisition of dataset, Feature extraction, Model training. Authors used a different mechanism than the transformer, which is pre-trained on about 21000 image datasets called ImageNet. To increase the model's accuracy, attention-based techniques have been developed to analyze the work effectively. After analyzing the attention mechanism, the transformer model is designed for efficient tasks (Chhabra et al., 2020). The machine that have employed in our proposed model constitutes a RAM of 8GB, AMD Ryzen 7 4700U with Radeon Graphics. Authors compared the SVM, Random Forest, KNN, Naïve Byes, etc., with the proposed approach to demonstrate the results more precisely.

3.1 Vision Transformer (ViT)

ViT was developed by Google and trained on approximately 300 million images through the dataset given by JFT (Gupta et al., 2019). Vision Transformer (ViT) approach has been proposed to identify skin disease. ViT is an upgraded variant of Transformer which requires linear information to perform accurately. This model splits the images into small patches, which are considered sequence tokens, and afterward, patches flattening takes place. It generates low-dimension embeddings linearly with the flattened patches to keep relevant information, and finally, the sequence output is passed as an input to the Transformer encoder. Without using convolution layers, it transforms the image into small-sized patches to avoid the difficulties with other state-of-the-art algorithms.

3.1.1 Vision Transformer over CNN

In CNN, the image which consists of pixels is dependent on the other neighboring pixels. For feature extraction, CNN makes use of filters on image patches (Gupta et al., 2019). CNN trains the model on the

selected features rather than considering all the features to make the model more precise. The drawback of CNN is that it does not care about the relative position of the features and cannot encode the spatial instruction. On the other hand, Vision Transformer trains the model, such as relative feature position, encoding spatial instruction, making it perform better than different state-of-the-art algorithms.

3.1.2 Working of the Vision Transformer on the dataset

Firstly, it divides the image into small-sized image patches. Tokens produced in the NLP application are the same as these image patches. Secondly, flattening the images patches is done and then resizing them into the vector format, which is passed as an input to the Transformer. After the vectorization of images, it converts into the embeddings linearly. Simultaneously, position embedding is passed through each of the embedding patches to preserve its positional information of the network. A class token like in BERT's method is integrated with the initial phase of the sequence of the embedded patches to make a learnable embedding.The Transformer encoder is the pre-eminent architecture comprising numerous encoder blocks, and every block has multiple self-attention mechanisms. The normalization of each layer is passed as an input to the multiheaded self-attention tool and Multilayer Perceptron (MLP) blocks (Mease et al., 2021).

3.1.3 Mathematical Intuition

The input sequence of 1D is pre-requisite as the input of the Transformer. For a 2D image, the image $W \in Q^{G \times V \times B}$ is transformed into flattened image patches $W_o \in Q^{m \times (O \times O \times B),}$ where (G, W) is the size of the original image, B is the channel count, (O, O) is the x, y coordinates of the image patches. $M = GWO^{-2}$ results in the required number of patches. Flattening of image patches is done, and it converts into vectors by projecting linearly. These are employed in a definite dimension. Then, the positional embedding of the patches is finished and associates the tokens, which act as an input to the transformer encoder.

There are two major components of the encoder block, i.e., Multiheaded self-attention layers and Multi-layer Perceptron Layer (MLP). Self-attention is also known as scalar dot-product attention. At first, calculate a query set $P = WV_P$, and keys set $J = WV_J$, and a set of values $U = WV_U$, where $W \in Q^{(M+1)xC}$ is the sequence input, $V_P \in Q^{C \times c}$, $V_Q \in Q^{C \times c}$, and $V_U \in Q^{C \times c}$ are the weights which can train, where c is the vector size. The self-attention mechanism is computed as:

Attention (P, J, U) = softmax $\left(\frac{PJ}{\sqrt{c}}\right)^{S}$ U (1)

Every row matrix is passed through the softmax function. In Multi-head Self Attention, j heads are appended to self-attention as follows:

MSA (P, J, U) = join ($head_0$, ……., $head_{j-1)}$ V (2)

$Head_{j-1}$ = Attention (Pj_{-1}, J_{j-1}, U_{j-1}) (3)

Each head deploys a sequence of size (0 + 1) x c. These j sequences are identified into (0 + 1) x cj sequence and MLP reorganized into (0 + 1) x c. The transformer encoder treats the image is given as follows:

$x_0 = [\, w_{class}; \mathrm{MLP}\left(w_o^1\right); \ldots; \mathrm{MLP}\left(w_o^0\right)] + D_{pos}$ (4)

$x'_m = \mathrm{MSA}(\mathrm{Norm}(x_{l-1})) + x_{l-1}$ (5)

$x_m = \mathrm{MLP}(\mathrm{Norm}(x'_m)) + x'_m$ (6)

$f = \mathrm{Norm}\left(w_M^0\right)$ (7)

The mathematical calculation of Vision Transformer has stated in the above Equations [1–7], how the model recognises the image, divides it into patches, and then flattens it. After flattening the patches, the image is transformed into a dimensional vector that passes as an input sequence to the Transformer encoder. Our proposed work to identify Melanoma skin disease by using ViT has been demonstrated by Fig. 2.

4. Results And Discussion

Here, the final obtained results by the Vision Transformer approach has been discussed. We've considered four frameworks to analyze the entire scenario as follows:

In scenario 1, authors proposed VGG-19 architecture that comprises 16 convolution layers, three fully connected layers, five max pool layers, and one softmax layer. 224*224 RGB images act as an input where 3*3 kernel size is used. It allows the dimensional padding to build the image's resolution, and subsequently, rectified linear unit is accessible for the improvement of less computation. Furthermore, the accuracy achieved by the VGG-19 is shown in Table 2.

In scenario 2, the Inception-V3 model has been conferred, which works completely on the convolution that splits n*n convolution kernel into 1*n and n*1. These are responsible for decreasing the number of parameters in training (Zhang et al., 2021). The main drawback of Inception-V3 is the absence of the residual association in the network to accomplish more precision in image classification. Using this deep neural network, can be achieved an accuracy of 86.59% with logistic regression, 87.31% with SVM, etc.

In scenario 3, authors analysed an experiment with a more efficient approach Vision Transformer that makes use of the self-attention mechanism, which is an efficient method to update the parallel input text to the embedding. From the critical examination of Table 2, it is observed that the highest accuracy has been achieved with ViT. We get the top-5 accuracy of 99% on our test data and the second-highest accuracy achieved by Inception-V3 as detailed in the following Table.

Table 2
Comparative study of various models in terms of different parameters using several approach

Feature extractor	Models	AUC	Accuracy	Precision
VGG-19 (Xiao et al., 2020)	KNNTree	0.90	0.81	0.81
	Tree	0.72	0.77	0.77
	Stack	0.93	0.85	0.85
	SVM	0.91	0.83	0.83
	Random Forest	0.91	0.83	0.83
	Neural Network	0.93	0.86	0.86
	Naive Bayes	0.82	0.76	0.76
	Logistic Regression	0.82	0.76	0.76
	CN2 rule inducer	0.75	0.67	0.67
Inception-V3 (Szegedy et al., 2016)	KNN	0.92	0.85	0.85
	Tree	0.74	0.75	0.75
	Stack	0.94	0.87	0.87
	SVM	0.94	0.87	0.87
	Random Forest	0.89	0.81	0.81
	Neural Network	0.95	0.89	0.87
	Naive Bayes	0.84	0.79	0.79
	Logistic Regression	0.95	0.86	0.86
	CN2 rule inducer	0.79	0.72	0.72
Transformer	Vision Transformer	0.98	0.99	0.99

In comparison with the other state-of-the-art algorithms, authors analyzed that the results shown by the ViT are magnificent. In CNN, the neurons perform the subtle detail (Nijhawan et al., 2017), but the other models failed to observe the proper orientation of the original image, which makes them less accurate. The relative positional information is not observed with different CNN, making them miss the relevant information. Therefore, we used the data augmentation technique for the proposed approach by transforming the image into proper orientation and trained the model after analyzing it by using the same method described in this paper. The Binary classification of Melanoma disease is shown in the Fig. 3.

This shows the accuracy of the proposed approach in context of different CNN architectures VGG-19, VGG-16, and Inception-V3 using bar graph. Various machine learning algorithms, SVM, KNN, logistic

regression, Random Forest, Ada boost, etc., are used in advance CNN. Y-axis shows the accuracy infraction.

5. Why Vision Transformer Is Better Than Cnn?

ViT uses a self-attention mechanism that allows the integration of the information entirely. This approach takes less time in training. It does not comprise a convolution layer, but it can handle the variable-sized information that makes it highly preferable. Implementing the self-attention technique on the image patches is a sophisticated task without CNN. ViT resolves this drawback by dividing the image into patches of size 16 by16. The Transformer encoder consists of Multi-Layer Perceptron (MLP) and Multi-head Self Attention Layer (MSA). The MSA divides the input into various heads to make the learning of each head easily. The result is associated and passed to the MLP. CNN functions on methods where pixels constitute the image where every pixel is dependent on the neighboring pixels. ViT allows the positional embeddings of the image while CNN only fetches the relevant information and features for training (Mondal et al., 2021). CNN fails to encode the relative spatial instruction.

6. Conclusion

This paper aims to determine the accurate prediction of the melanoma skin cancer and to validation of the obtained results compare it with the other CNN models. From the comparative study, it is concluded that our proposed method provides the highest accurate results than VGG-16 and Inception-V3. It is observed that authors have achieved top-5 accuracy of 99% which is better than the other described state-of-the-art algorithms. So, ViT is found to be more accurate than other CNN models. In addition to the model, we also used data augmentation to enhance the classification accuracy.

This paper is very effective for patients and dermatologists to treat skin cancer more precisely. Furthermore, the application of our paper can be very helpful for the early detection of Melanoma to take precautions. Our work can also be of great usage to a dermatologist and medical experts to identify the disease with limited manpower.

Declarations

Ethical Approval

There is no humans or animals related research are involved in this work.

Competing interests

Authors have no conflict of interest.

Authors' contributions

All the authors contributed equally in this work under the mentorship of Dr Nupur Goyal.

Funding

No funding is required for this.

Availability of data and materials

Data used in this work is taken from Kaggle

References

1. Vijayalakshmi, M. M. (2019). Melanoma skin cancer detection using image processing and machine learning. International Journal of Trend in Scientific Research and Development (IJTSRD), *3*(4), 780–784.
2. Mengistu, A. D., & Alemayehu, D. M. (2015). Computer vision for skin cancer diagnosis and recognition using RBF and SOM. International Journal of Image Processing (IJIP), *9*(6), 311–319.
3. Cıcero, F., Oliveira, A., Botelho, G., & da Computacao, C. D. C. (2016). Deep learning and convolutional neural networks in the aid of the classification of melanoma. In *Proc. SIBGRAPI* (pp. 1–4).
4. Garnavi, R., Aldeen, M., & Bailey, J. (2012). Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis. IEEE transactions on information technology in biomedicine, *16*(6), 1239–1252.
5. Celebi, M. E., Iyatomi, H., Schaefer, G., & Stoecker, W. V. (2009). Lesion border detection in dermoscopy images. Computerized medical imaging and graphics, *33*(2), 148–153.
6. Garnavi, R., Aldeen, M., Celebi, M. E., Bhuiyan, A., Dolianitis, C., & Varigos, G. (2010). Automatic segmentation of dermoscopy images using histogram thresholding on optimal color channels. International Journal of Medicine and Medical Sciences, *1*(2), 126–134.
7. Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. Computerized Medical imaging and graphics, *31*(6), 362–373.
8. Stolz, W. J. E. J. D. (1994). ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. Eur. J. Dermatol., *4*, 521–527.
9. Ng, V. T., Fung, B. Y., & Lee, T. K. (2005). Determining the asymmetry of skin lesion with fuzzy borders. Computers in biology and medicine, *35*(2), 103–120.
10. Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics, *34*(1), 28–36.
11. Gilmore, S., Hofmann-Wellenhof, R., & Soyer, H. P. (2010). A support vector machine for decision support in melanoma recognition. Experimental dermatology, *19*(9), 830–835.
12. Menzies, S. W., Bischof, L., Talbot, H., Gutenev, A., Avramidis, M., Wong, L., Lo, S. K., Mackellar, G., Skladnev, V., McCarthy, W., Kelly, J., Cranney, B., Lye, P., Rabinovitz, H., Oliviero, M., Blum, A., Virol, A.,

De'Ambrosis, B., McCleod, R., Koga, H., Grin, C., Barun, R. & Johr, R. (2005). The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. Archives of dermatology, *141*(11), 1388–1396.

13. Manousaki, A. G., Manios, A. G., Tsompanaki, E. I., Panayiotides, J. G., Tsiftsis, D. D., Kostaki, A. K., & Tosca, A. D. (2006). A simple digital image processing system to aid in melanoma diagnosis in an everyday melanocytic skin lesion unit. A preliminary report. International journal of dermatology, *45*(4), 402–410.
14. Rawat, S. S., Bisht, A., & Nijhawan, R. (2019, November). A Deep Learning based CNN framework approach for Plankton Classification. In *2019 Fifth International Conference on Image Information Processing (ICIIP)* (pp. 268–273). IEEE.
15. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.
16. Chhabra, H. S., Srivastava, A. K., & Nijhawan, R. (2020). A hybrid deep learning approach for automatic fish classification. In *Proceedings of ICETIT 2019* (pp. 427–436). Springer, Cham.
17. Gupta, S., Panwar, A., Goel, S., Mittal, A., Nijhawan, R., & Singh, A. K. (2019, December). Classification of lesions in retinal fundus images for diabetic retinopathy using transfer learning. In *2019 International Conference on Information Technology (ICIT)* (pp. 342–347). IEEE.
18. Mease, P. J., Liu, M., Rebello, S., McLean, R. R., Dube, B., Glynn, M., Hur, P. & Ogdie, A. (2021). Association of nail psoriasis with disease activity measures and impact in psoriatic arthritis: Data from the CORRONA psoriatic arthritis/spondyloarthritis registry. The Journal of rheumatology, *48*(4), 520–526.
19. Zhang, M., Su, H., & Wen, J. (2021). Classification of flower image based on attention mechanism and multi-loss attention network. Computer Communications, *179*, 307–317.
20. Xiao, J., Wang, J., Cao, S., & Li, B. (2020, April). Application of a novel and improved VGG-19 network in the detection of workers wearing masks. In *Journal of Physics: Conference Series* (Vol. 1518, No. 1, p. 012041). IOP Publishing.
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).
22. Nijhawan, R., Sharma, H., Sahni, H., & Batra, A. (2017, December). A deep learning hybrid CNN framework approach for vegetation cover mapping using deep features. In *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 192–196). IEEE.
23. Mondal, A. K., Bhattacharjee, A., Singla, P., & Prathosh, A. P. (2021). xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography. IEEE Journal of Translational Engineering in Health and Medicine, *10*, 1–10.

Figures

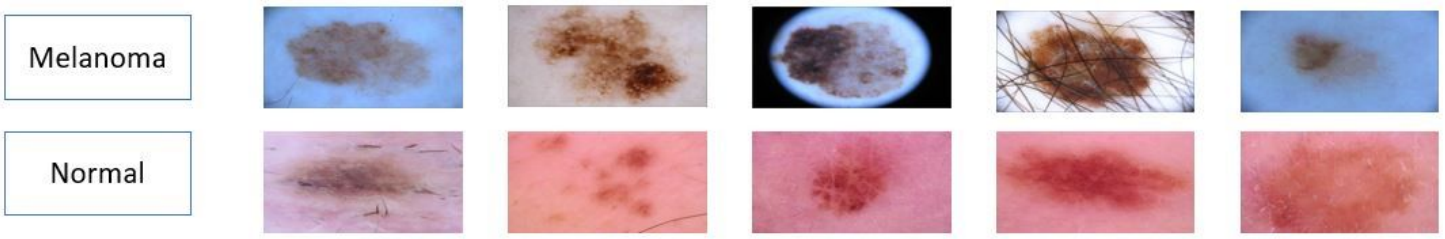


Figure 1

Sample images of skin from our dataset

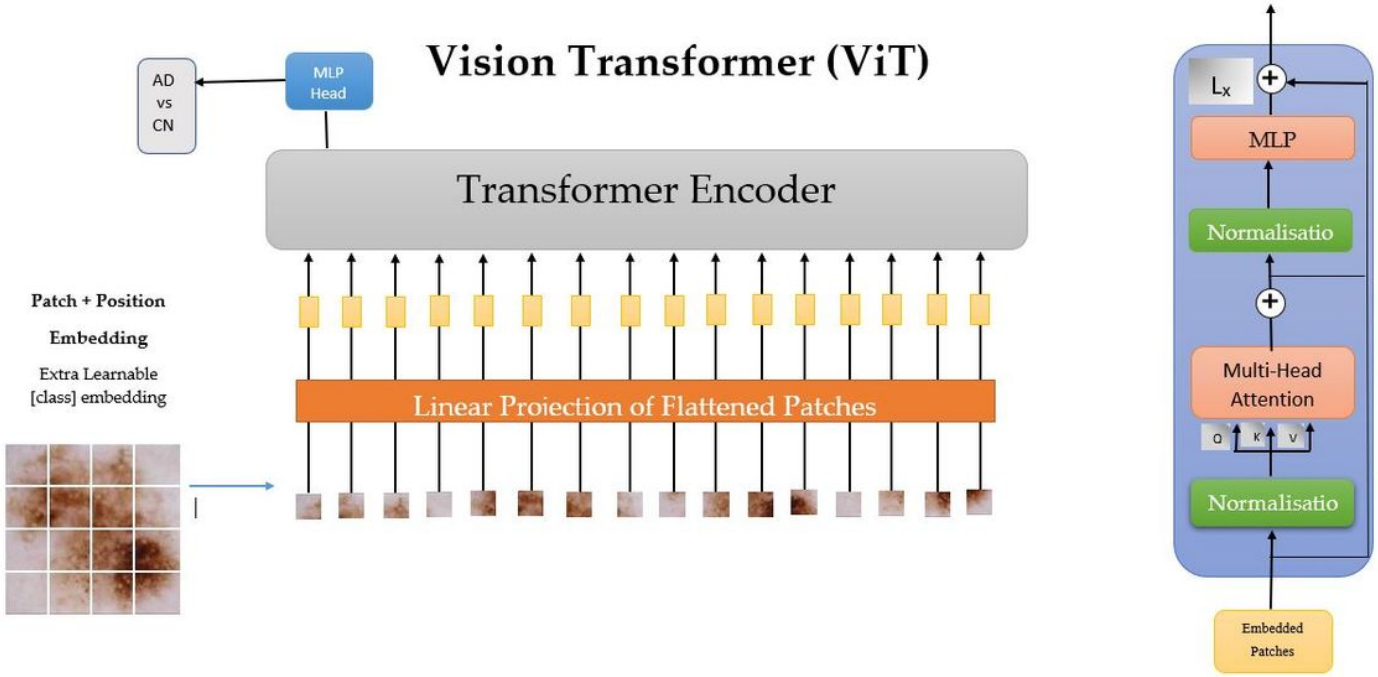
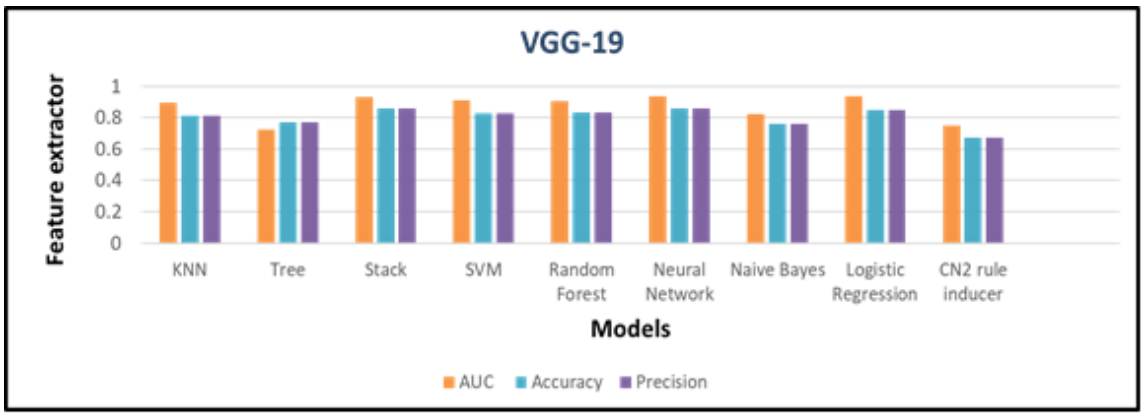
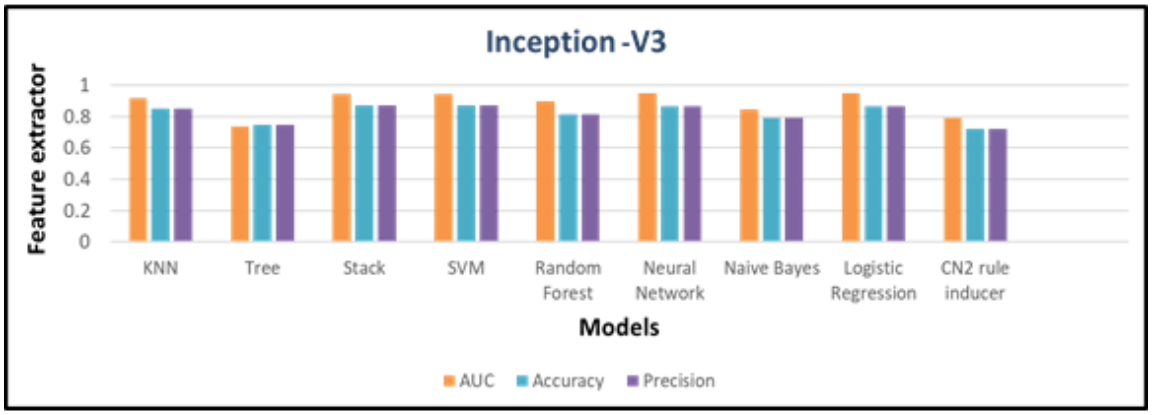


Figure 2

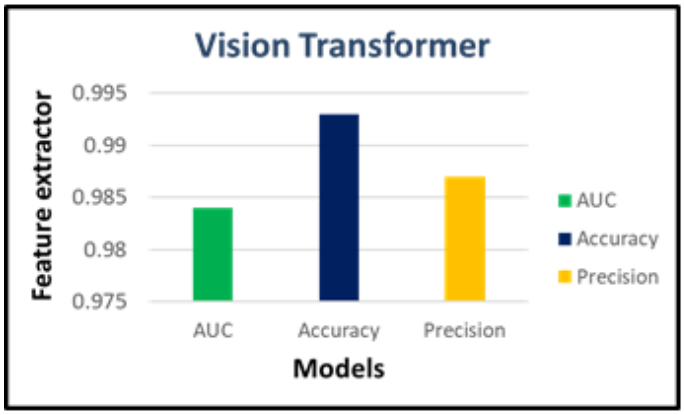
Architecture of vision transformer



(a)



(b)



(c)

Figure 3

Binary classification results of Melanoma disease, (a) VGG-19, (b)Inception-V3 and (c) Vision Transformer