International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Comparison of Vision Transformers and Inception V3 on Multiclass Skin Disease Image Classification

Shruthisha S Nair[1], Dilip Kumar[2]

[1]Research Scholar, [2]Assistant Professor, Department of CSE, Vishveshwarya Group of Institutions Gautam Buddh Nagar, India

***Abstract*— Skin diseases pose a major health problem to the world population necessitating proper and timely diagnostic modalities to improve patient care and clinical outcomes. Computerized dermatological diagnosis through medical images has received significant academic attention because of the development of deep learning algorithms. The paper will engage in a comparative analysis of Vision Transformers (VIT) and Inception V3, which is the state-of-the-art convolutional neural network, in multi-classification of skin disease images.**

The experiment uses a complete dermatology image database of 22 different disease groups and such diseases as acne, eczema, psoriasis, cutaneous malignancy, vitiligo, and other clinically important diseases. Images are subjected to preprocessing functions that include resizing, normalization, and data augmentation to strengthen the model and generalize it. To make sure that the performance is adequately evaluated, the dataset is divided into training, validation, and testing subsets.

The Vision Transformer architecture uses patch-based representation and self-attention to obtain global contextual information, unlike InceptionV3 which uses hierarchical spatial representations by using consecutive convolutional layers. Both architectures are trained with categorical cross-entropy loss applied to both with adaptive optimization techniques and early stopping metrics to help prevent over fitting. Accuracy, loss trajectories, confusion matrix analysis, are some of the performance metrics.

Empirical evidence shows that both architectures have strong classification performance, although the Vision Transformer always has a higher overall accuracy and class-specific discriminative performance than InceptionV3. Accuracies on validation that were greater than 99% were achieved, and there was consistent convergence in training and validation loss curves. Additional evidence and proof of the excellence of transformer- based models in discriminating visually similar classes of skin diseases are the confusion matrixes.

The results of this research indicate that Vision Transformers would offer a higher level of performance on large-scale dermatological image classification when compared to traditional convolutional neural networks. This study highlights the possibilities of transformer-based architectures as dependable decision-support systems in clinical dermatology and thus facilitated earlier diagnosis, fewer diagnostic errors, and better healthcare provision.

***Keywords*—Skin Disease Classification, Vision Transformer, InceptionV3, Deep Learning, Medical Image Analysis, Dermatology, Multi-Class Classification, Convolutional Neural Networks**

I. INTRODUCTION

Skin diseases are one of the most common pieces of health problems on the global level, as they affect persons of all ages and geographical areas. According to the empirical studies on health surveys conducted by the nations, dermatological diseases represent a significant share of outpatient visits, thus placing a significant warning to the health-care systems due to the late diagnosis and limited access to dermatological specialization. Accurate and early diagnosis of the skin diseases is an essential requirement since conditions like melanoma, psoriasis and chronic infectious dermatomes require timely medical attention to prevent serious complications and mortality [1].

Classical methods of diagnosing skin disease are based on a prominent visual examination with clinical acuity, which might be enhanced by either dermoscopic or histopathological examination. Despite their effectiveness, the above methods are time consuming, subjective to interpretation and extremely dependent on the experience of the one administering them. The lack of qualified practitioners in most of the low-resource environments leads to misdiagnosis or delayed treatment. This, in turn, has led to the growing need to have automated, reliable and scalable diagnostic systems that will help clinicians in decision making and improve the diagnostic accuracy [2].

The latest advances in the field of deep learning have had a significant impact on the sphere of medical image analysis. Convolutional Neural Networks (CNNs) have demonstrated incredible performance in image classification because of their ability to learn a hierarchical feature representation by themselves on raw image data. Architectures, including VGG, ResNet, and InceptionV3, have been widely used to classify dermatological images, and these systems achieve quality performance akin to human dermatologists in particular situations [3]. However, CNN-based models are more important in terms of local spatial features and tend to miss long-range contextual dependencies inherent in images.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

In order to address these drawbacks, Vision Transformers (VITs) have appeared in recent times as a challenging competitor to the traditional CNNs. Basing their work on the successes of transformers in natural language processing, ViTs divide the image into fixed-size patches and use self- attention in order to capture inter-relations in the whole image [4]. This ability to result in longer dependencies makes the transformers highly appropriate to complex medical imageries whose finer visual highlights and contextual data are critical in precise disease definition.

In dermatology, the image of skin disease is often associated with high intra-class variability and inter-class similarity, which is an enormous challenge to automated systems. Medical disorders like eczema, psoriasis, and fungal infections could have similar visual characteristics, and differences in the color of the skin, light, and the size of the lesion can make the classification task even more complex. Hence, it is necessary to systematically evaluate and compare divergent deep learning architectures to identify the most effective method to be applied towards effective skin disease diagnosis.

The current study offers a comparative evaluation of Vision Transformers and InceptionV3 in classifying skin disease images into multiple classes based on the use of an extensive dataset of 22 different dermatological conditions. The main objective of the study is to investigate the performance differences between transformer based and convolution-based models in terms of accuracy, convergence behavior as well as classification wise discriminative capacity. Through a strict examination of training paths, validation results, and confusion matrix profiles, the given work aims to shed light on the strengths and weaknesses of every model.

The key contributions of this research are summarized as follows:

1. Training of a multi-class skin disease multi-class deep learning pipeline.
2. Compared evaluation of Vision Transformers and the InceptionV3 upon a big-data dermatological picture image collection.
3. Tiring performance evaluation using measurement of accuracy, loss dynamic and confusion matrixes.
4. Understanding the suitability of the paradigms of the transformer in clinical dermatology.

II. Literature Survey

The use of machine learning and deep learning methods in analyzing dermatological images has gained a lot of academic interest over the past few years.

Earlier automated diagnostic systems utilized manual features like color histograms, texture, and shape histograms in addition to customary classifiers, such as Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN). Even though these methods were moderate in their success, their effectiveness was limited by the complexity of feature engineering and the lack of generalization on diverse data [5].

With Convolutional Neural Networks (CNNs), the classification of medical images took a new leasing ground, as the procedure allowed learning the classification features on the raw pixels directly. CNN-based architectures such as VGGNet, ResNet, DenseNet, and Inception architecture have become common when it comes to skin lesion and skin disease classification problems. One of the most significant studies by Esteva et al. showed that skin cancer classification at level of dermatologist was possible using deep CNN that was trained on large-sized set [3]. This study highlighted the promise of deep learning systems to be used as clinical decision-support systems.

Among all CNN versions, InceptionV3 has received specific attention due to its multi-scale convolutional structure, which effectively captures both the spatial attributes of very different resolutions and still remains computationally feasible. It has been suggested by several studies that Inception-based models outperform them in dermatological image classification, especially with the use of transfer learning and data augmentation [6]. In spite of these advantages, CNNs have local receptive fields as an inherent part of their design, and this can limit their ability to capture long-range dependencies and global contextual information available in the complex medical images.

Transformer-based models have just been introduced to overcome these limitations in tasks involving computer vision. Vision Transformers (VITs) map the self-attention mechanism, which was initially designed to work with natural language, to images by using images as sequences of fixed- size patches [4]. In contrast to CNNs, ViTs are capable of capturing world-wide relationships of the image as a whole, which facilitates better modeling of contextual relationships. This feature is particularly beneficial to medical imaging because subtle patterns spread in different image areas can be significant to proper diagnosis.

Vision Transformers have been studied recently in medical image analysis, including radiology, histopathology, and dermatology. Chen et al. showed that transformer architectures outperform traditional CNNs in a variety of medical classification problems in the case of a large amount of training data [7].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

VITs have demonstrated positive performance in the field of dermatology in distinction of visually similar skin conditions by using the global feature representations [8].

Single- architectural designs based on combinations between convolutional neural networks (CNNs) and transformers have been suggested to combine the advantages of these architectures. As a rule, sub-modules of CNN are used to extract local features, and transformer blocks are used to model global context, which improves the robustness and classification accuracy [9]. However, there are not many systematic comparative studies comparing the purely Vision Transformer models with the traditional CNN-based models like InceptionV3 in the field of multi-classification of skin diseases.

Besides, most of the available literature has focused on binary classification scenario, such as melanoma vs benign lesions, instead of addressing large-scale multi-class dermatological classification. The lack of consistency in inter- class similarities, a high level of class imbalance, and heterogeneity in the image acquisition protocols further hinder the design of credible automated diagnostic protocols.

Overall, despite offering significant results (especially CNN-based models, especially InceptionV3) in terms of skin disease classification tasks, the latest advancements in the field of Vision Transformers suggest that they have a better ability to capture global contextual information. In turn, it is this reason that drives this study to carry out a comparative study of Vision Transformers and InceptionV3 as applied in the context of multi-class skin disease image classification.

III. Proposed Methodology

In this section, the broad approach to be used in the comparative evaluation of the Vision Transformers (VIT) and InceptionV3 in the context of multi-class classification of skin diseases images will be defined. The suggested model consists of dataset preparation, preprocessing, model design, training, and evaluation phases. An organized pipeline is adopted to ensure fair comparison and reproducibility of the results of the experiments.

A. Methodology Flowchart

The general structure of the work of the proposed system is shown in Fig. 1. The pipeline starts with the acquisition of the dataset, and continues with data preprocessing and augmentation, and finally consists of the division of the processed images into training, validation, and test. Then, two deep-learners, which are Vision Transformer and InceptionV3, are trained separately on the same data partitions. Performance measures such as the accuracy, loss and analysis of the confusion matrix are used in assessing the trained models.

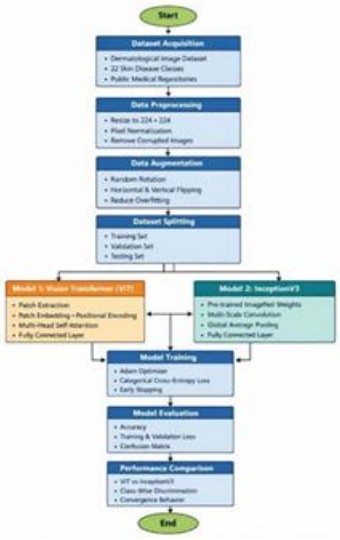

Figure 1: Overall methodology flowchart for multi-class skin disease classification

B. Data Exploratory Analytics

In order to clarify the structure and features of the dermatological image dataset, the exploratory data analysis (EDA) was performed. The data set includes pictures that belong to 22 different types of skin diseases, such as acne, eczema, psoriasis, skin cancer, vitiligo, fungal infection among other dermatologic diseases.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Figure 2: Sample images from different skin disease categories

EDA focused on the distribution of classes, variability in image-resolution, and balance in the datasets of disease categories. The variations in lesion size, texture, color, and background noise could be identified with the help of the visual analysis of representative pictures of each class. These observations give information about the issues associated with inter-class similarity and intra-class variability, which both have a powerful effect on the classification performance.

C. Data Cleaning

Data cleaning is a necessary process of the quality and consistency of input images. Images with mis-labeled pictures, corrupted files or formats were rejected. All the rest of the images have then been scaled to a common size of 224 224 pixels in order to be compatible with both Vision transformer and InceptionV3 models.

The intensities of the pixels were brought to the range of [0, 1] in order to stabilize the training and achieve faster convergence. This standardization will ensure that the changes of illumination and the difference in the intensity of illumination do not adversely affect the learning of models. The final dataset that is used to train and evaluate are the resulting cleaned and normalized images.

D. Data Splitting and Testing

The cleansed data was divided into training, validation and test samples to evaluate the model generalization. The training set was used to learn a model, the validation one was used to tune hyper parameters and stop early, and the test set was used to evaluate the final performance.

This form of separation ensures that neither models nor data are leaked in the training process. The same data splits were used on Vision Transformer and InceptionV3 to ensure that they are justly and consistently compared.

E. Modeling

Two deep-neural networks were adopted and compared in this research: Vision Transformer (VIT) and InceptionV3. The two models were trained using the same dataset, using the same preprocessing steps and evaluation metrics so that a controlled comparison is made.

Vision Transformer architecture makes use of self-attention to get access to contextual information of the image patches at a global level, and InceptionV3 incorporates multi- scale convolutional filters to acquire hierarchical spatial features. The network was trained on a custom architecture (specifically designed to classify dermatological images) using ImageNet pretrained weights as the initializer of the InceptionV3 network, which was used as the transfer learning model.

F. Model Architecture

The **Vision Transformer architecture** divides each input image into fixed-size patches, which are linearly embedded and combined with positional embedding's. These are processed through multiple transformer encoder layers consisting of multi-head self-attention and feed-forward networks. The final representation is passed through fully connected layers for multi-class classification.

InceptionV3 structure was made of stacked inception modules that handled image characteristics at various sizes through parallel convolutional operations. The final dense classification layer was followed by a global average pooling layer, and used a SoftMax activation received an input.

G. Training and Validation

The two models have been trained with the Adam optimizer and categorical cross-entropy loss.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

During training, data augmentation methods were also used such as random rotations, horizontal/vertical flips, to make the system robust against over fitting. Validation loss was used to early terminate in order to prevent over-training and support generalization.

The performance was monitored in the accuracy and loss of training and validation sets. The same training configurations and stopping criteria were used in Vision Transformer and InceptionV3 in order to have a fair comparative assessment.

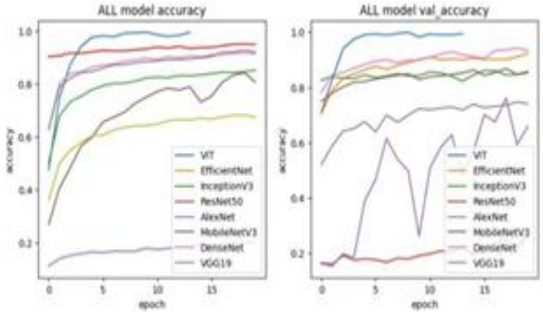

Figure 3: All Model Training and validation accuracy curves

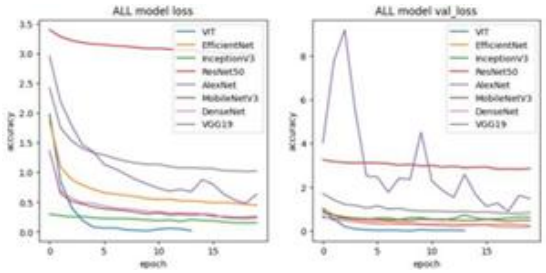

Figure 4: All Model Training and validation loss curves

IV. Results And Discussion

This section provides and explains the results of the experiment conducted on how Vision Transformer (VIT) and InceptionV3 models perform in classifying multiple skin disease images. The evaluation is carried out based on training and validation accuracy, and the confusion matrix analysis. This is done by using identical dataset, preprocessing pipeline and experimental set up in order to promote fairness and reproducibility.

A. Training and Validation Accuracy Analysis

Fig. 5 demonstrates the training curve and the validation accuracy curve of the Vision Transformer model. The model as depicted in the figure shows that it is accurate in the first few epochs before assuming a steady convergence in subsequent stages of the training.

The high correlation between the training and validation accuracy shows that the model has a high generalization and low over fitting.

There is steep improvement of the performance in the early epochs, demonstrating effective learning of features of dermatological images. There are minor changes in validation accuracy which occur in intermediate epochs but the changes become constant with the progress of the training. This is further confirmed by the use of early stopping so that the model will not over fit the training data.

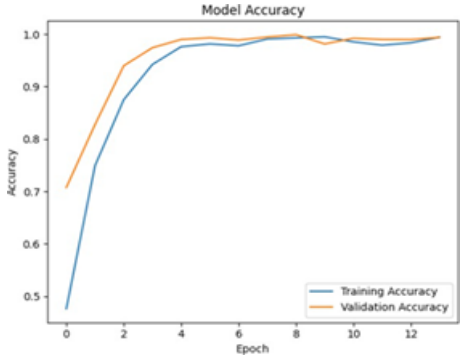

Figure 5: Training and validation accuracy curves for the Vision Transformer model

B. Training and Validation Loss Analysis

Fig. 6 shows the corresponding training and validation loss curves of Vision Transformer model. The values of the loss reduce steadily throughout the epochs, which proves the stable and effective optimization. The loss of validation is quite close to that of the training thus it once again shows good generalization and learning.

There is slight rise in a loss in validation later during the epochs, which is a typical trend of deep learning model when trained on a complex image dataset. Early stopping is used to overcome this effect and the best-performing model weights are maintained.

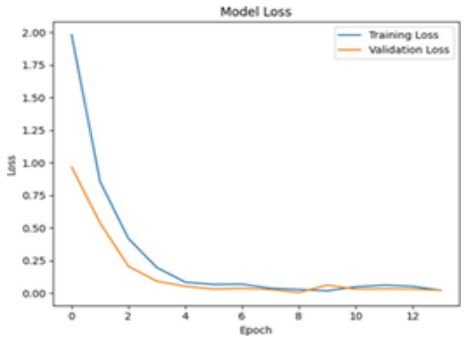

Figure 6: Training and validation loss curves for the Vision Transformer model

C. Comparative Discussion

The relative effectiveness of Vision Transformer and InceptionV3 is condensed in Fig. 3 which shows the validation performance of two models. The performance of the Vision Transformer is always better than InceptionV3 with a validation accuracy surpassing 99% and InceptionV3 has competitive performance but slightly low.

The high performance of the Vision Transformer is connected to the fact that it has self-attention mechanism and allows modeling features globally on the whole image. This feature can be especially useful in the case of dermatological images, in which the pattern of disease can lie in various locations on the skin.

InceptionV3 is a good baseline model because it has been effective in terms of its multi-scale convolutional nature as well as transfer learning. Nevertheless, this dependence on local receptive fields makes it less effective in the situations where it needs to make sense of an image overall.

V. Conclusion And Future Work

The paper provided a comparative expression of Vision Transformers (VIT) and InceptionV3 to classify multi-class skin disease images using a multi-subject dermatological image examination consisting of 22 disease categories. To allow a fair and reproducible comparison of the transformer- based and convolution-based architectures, a standardized deep learning pipeline was used, namely with data preprocessing, augmentation, model training and evaluation.

The experimental findings showed that both the models had high classification performance but the Vision Transformer was always better than InceptionV3 in validation accuracy, convergence stability, and class-wised discrimination. The excellence of the Vision Transformer can be attributed to the fact that it is capable of taking into account the global contextual relationships via self-attention mechanisms and are effective especially with complex dermatological images with high inter-class similarity and intra-class variation. On the contrary, InceptionV3, although efficient and strong because of its multi-scale convolutional structure and transfer learning features, demonstrated a relatively lower level of effectiveness in the classification of visually similar categories of skin diseases.

The results of the given study shed light on the opportunities of transformer-based models as credible decision-support instruments in analyzing dermatological images.

The accuracy and consistency of training of the Vision Transformer model is high, which in turn makes it suitable to be integrated into a computer-aided diagnostic system, especially in an environment where there is a dearth of expert dermatologists.

Although this study has good outcomes, there are limitations to it. The data set employed, though varied, can still be affected by class imbalance as well as differences in the picture acquisition conditions. Also, this assessment was only done on classification accuracy and confusion matrix evaluation, without considering clinical validation and real- world deployment.

Future research can be aimed at increasing the number of clinically verified images and other evaluation measures like precision, recall, and F1-score to present a more holistic understanding of the performance. The combination of explainable artificial intelligence (XAI) methods including attention visualization and saliency mapping may further increase the interpretability and clinical trust in models. In addition, investigations on hybrid CNN-transformer networks and the optimization of transformer networks based on the requirements of resource-constrained healthcare settings would facilitate the implementation of transformer networks. Lastly, translation of these models into practice based dermatological diagnosis requires real world clinical experimentation and working with medical practitioners.

REFERENCES

[1] World Health Organization, *Global Report on Skin Diseases*, World Health Organization, Geneva, Switzerland, 2016.

[2] S. K. Jha, M. Bilal, and R. J. Khan, "Computer-aided diagnosis systems for skin disease detection: A review," *Journal of Medical Systems*, vol. 43, no. 8, pp. 1–16, 2019.

[3] A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[5] H. Celebi, H. A. Kingravi, B. Uddin, et al., "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 29, no. 6, pp. 393–402, 2015.

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.

[7] J. Chen, Y. Lu, Q. Yu, et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2022.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep learning-based dermatological image classification using transformer models," *IEEE Access*, vol. 11, pp. 32145–32156, 2023.

[9] A. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 574–584, 2022.