2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) | 979-8-3503-5531-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/EECSI63442.2024.10776508

Performance Assessment of Advanced CNN and Transformer Architectures in Skin Cancer Detection

Md Saiful Islam Sajol
Dept. of CSE
Louisiana State University
Baton Rouge, USA
msajol1@lsu.edu

Syada Tasmia Alvi
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
syada.cse@diu.edu.bd

Chowdhury Abida Anjum Era
Dept. of Computer Science
Fachhochschule Dortmund
Dortmund, Germany
chowdhury.era001@stud.fh-dortmund.de

***Abstract*—Skin cancer has always been one of the most common types of cancer in medical history. Every year, it kills millions of people around the world. To find skin problems early, it is necessary to have a reliable automated method for recognizing them. In the past, protein sequences and different types of imaging methods were used with machine learning to find skin cancer. The problem with machine learning methods is that they need features to be designed by humans, which is hard to do and takes a lot of time. Image processing and deep learning are both used in the method for treating skin cancer that works. Our study is based on The HAM10000 dataset, which is made up of 10015 different dermatoscopic pictures of common pigmented skin lesions from different sources. The newest deep learning techniques, such as Convolutional Neural Networks (CNN), Transformers, and Hybrid models, i.e., ConvNeXt V2, Swinv2, ViT2, Cvt, DenseNet, RestNet, PVTv2, EfficientNet, EfficientFormer, VGG, MobileNet, MobileViTV2, GoogLeNet, are all compared in this study to see which one is better for automatically detecting skin cancer lesions. We found the best model for identifying skin lesions by determining its accuracy, F1-score, Inference time, and confusion matrix. Among our implemented models, the ConvNeXt V2 model has an accuracy of 93.2% with the lowest inference time. With this research's help, new ways are being made to find skin cancer, which can lead to better patient results and better clinical decisions.**

***Index Terms*—Deep learning, Skin cancer, HAM10000, Convolutional neural network**

I. Introduction

Since the beginning of this decade, skin cancer has been the most common type of cancer. The skin is the biggest organ in the body because it covers most of it [1]. The human body may be shielded from UV radiation, heat, injuries, and bacterial or viral infections by its skin. The skin stores water and fat in addition to aiding in the regulation of body temperature. One of the most prevalent types of cancer and the biggest public health concern is skin cancer [2]. It kills more people than heart disease and affects more than a million people every year. So, by the time a person turns 70, one in five Americans will have skin cancer, and more than two people die every hour in the United States from skin cancer [3]. Mainly, uncontrollably growing DNA mutations are the primary cause of skin cancer [1]. The risk of skin cancer may increase if a person has sunburned more than five times [3]. In order to save time, effort, and lives, it is essential to have a reliable automatic method for identifying skin lesions as soon as possible [4].

Cost-effective cancer diagnosis can be achieved by computer-aided design. By using several imaging techniques, cancer detection may be evaluated. Evaluating and analyzing may be time-consuming and error-prone. Complex skin lesion images are the main reason. Machine learning algorithms have been used to classify skin lesions [5]. Specialists can quickly find cancer with the help of automatic skin lesion clarification. [5]. Human-engineered characteristics are necessary for machine learning algorithms [1]. Only in situations where there are experts available can machine learning be employed, and choosing the right features takes a great deal of effort. There will be data loss at the beginning of the preprocessing procedures, which could lower the classification quality. A bad segmentation outcome, when taking into account the sample example, often leads to a poor featrue extraction outcome and, thus, a low classification accuracy. The best method for identifying skin cancer is to use Deep Learning algorithms [5]. The standard procedure for identifying and diagnosing skin cancer involves obtaining the image, preprocessing it, segmenting the resulting preprocessed image, extracting the relevant features, and categorizing it. Various deep learning methods have been used for datasets in order to analyze, identify, and classify photographs [1]. Recently, convolutional neural networks (CNNs) and Transformer networks have been employed in learning respectively the local and global features of lesion images. In this work, we compare the skin cancer classification performance of CNN, transformers, and hybrid models. We preprocess the data first. Next, we fill our dataset with data augmentation techniques such as rotation angles and horizontal flips. Ultimately, we construct and refine the models so they can forecast. For transformers, the baseline models selected are ViT2, SwinV2, and PVT V2. ConvNeXt V2, VGG, GoogLeNet, DenseNet, ResNet, EfficientNet, MobileNet, and InceptionV3 are the models selected for CNN, while CvT, EfficientFormer, and MobileViTV2 are the models selected for hybrid (CNN+Transformer). With an accuracy of 92.66%, the ViT2 model is the most accurate transformer-based model. The ConvNeXtV2 model has the best accuracy among CNN models, at 93.2%. The accuracy of CVT in the hybrid model is 92.20%. In general, the CNN models outperformed the other models when it came to classifying skin cancer. Experiments on a famous dataset, the HAM-10000, have confirmed the effects of the proposed method in terms of improved accuracy score, recall, precision, F1 score, and inference time.

The remainder of the paper is structured as follows: section II presents work that is related, section III describes the proposed method, section IV presents the results and discussion, and section V reviews the conclusion and future work.

II. Related Works/Literature Review

Early detection of skin cancer is very important as a means of preventing the severity of the disease [6]. There has been an enormous amount of work done for skin cancer detection using machine learning and deep learning approaches.
In order to detect different types of skin cancer, Arshed et al. [7] suggested a different approach that relies on commercially available ViT. Eleven existing deep learning methods that employ convolutional neural networks (CNNs) to transmit data were compared to the suggested method to see how well it performed [8]. Yang et al. [9] introduced a new Transformer that can classify skin lesions. It was necessary to do rebalancing, preprocessing (flattening and dividing into patches), transformerencoding, normalization, and classification in order to reach the study's purpose. The results demonstrated an average of 94% accuracy, 82% precision, 84% sensitivity, 82.28% F1, and 97.55% specificity, using the HAM10000 skin cancer dataset. Azhar Imran et al. [2] utilized an ensemble technique to address skin cancer by combining the VGG, ResNet, and Caps Net models. They showed that the ensemble model outperformed the individual learners in decision-making and perception-related matters. The EfficientNet B4 model was proposed in [10] for the purpose of classifying skin cancer. The accuracy of this model reached 87.90%. In [11], Convolutional Neural Networks are primarily employed to identify and categorize seven distinct types of skin lesions. Initially, skin lesion photos are examined to determine their features. This is achieved by employing image processing technologies and sampling techniques to preprocess the images. Next, the training features of the imageNet network are fine-tuned using transfer learning techniques on InceptionV3, ResNet50, DenseNet201, and other networks for classification training. The skin lesion photos were classified using a deep convolutional network called DenseNet201. The network achieved an impressive classification accuracy of 99.12% for training and 86.91% for testing. Zenebe et al. [12] developed a promising attention-augmented ConvNeXt network for accurate skin lesion classification. The system uses channel and spatial attention processes to improve its capacity to focus on informative visual parts. A deep learning model based on Swin Transformer V2 was suggested by Li et al. [13] for speedy and accurate fundus disease diagnosis. The innovative Swin Transformer model excels in several computer vision applications. This approach reduced computing complexity and improved classification efficiency by estimating self-attention inside local windows. Zheng et al. [14] offer the PVTCOV19D, an automated COVID-19 diagnostic framework using lung CT scan images. PVT uses a pyramid structure in the transformer to perform dense prediction tasks like detection and segmentation. The approach is utilized for COVID-19 detection in this publication. Fraiwan and Faouri [15] achieved 82.9% accuracy using DenseNet201 on the HAM10000 dataset with seven classes. Multiple models, such as SqueezeNet, GoogLeNet, Inceptionv3, DenseNet-201, MobileNetv2, ResNet18, RestNet50, ResNet101, Xception, InceptionResNet, ShuffleNet, DarkNet-53, and EfficientNet-B0, were examined to determine accuracy. [16] [17] suggested an ensemble technique using Swin-Transformer and EfficientNetV2S to detect skin cancer's first focus region, making use of dark image regions. In [18], a CVT-Vnet is proposed for multi-organ segmentation in 3-dimensional CT scans of head and neck cancer patients for radiation treatment planning. Organs such as the brainstem, chiasm, mandible, optic nerve, parotid, and submandibular are present. Gururaj et al. [19] employed the MNIST: HAM10000 database for their experimental study. The pre-processing techniques utilized were autoencoder and decoder segmentation, dull razor, and sampling. The model was trained using transfer learning techniques, specifically DenseNet169 and ResNet50, to achieve the required results. The undersampling strategy employed in DenseNet169 attained an accuracy of 91.2% and a f1 measure of 91.7%. In contrast, the oversampling technique used in ResNet50 provided an accuracy of 83% and a f1 value of 84%. According to Ding et al. [20], HI-MViT is a lightweight model for skin disease classification that uses fewer parameters to represent local and global information for rapid and reliable dermoscopic image classification on mobile devices. Uses MobileNet29 and Vision Transformer hybrid architecture MobileViT, with the new Improved-MV2 block reducing computations and improving mobile terminal landing. In experiments on the ISIC-2018 dataset, F1-Score, Accuracy, AP, and AUC attained high scores of 0.931, 0.932, 0.961, and 0.977, respectively.

III. Research Method

The proposed system architecture for skin cancer detection is shown in Figure 2 as a sequence of phases. Cancerous pictures were pre-processed to improve quality. After pre-processing, the data was split into training and testing sets. Then, each model will be trained successively using train data. After system completion, testing data will be used to evaluate classification accuracy.

A. Collection of Data / Dataset Description

To train, validate, and test the model, one needs a dataset, which is a collection of relevant data. Because of its outsized effect on model performance, the dataset is a crucial component in and of itself [6]. HAM1000 data set is the set that was used in this method. Through the Harvard Dataverse [21], anyone can get to it. Human Against Machine has 10,015 pictures, which is what HAM10000 stands for. The sample has 10,015 dermoscopic images that are split into 7 groups. These come from different groups of people and are gathered and stored in different ways. The information was taken straight from Kaggle. The skin lesions that are looked at for classification are Actinic keratoses and intraepithelial

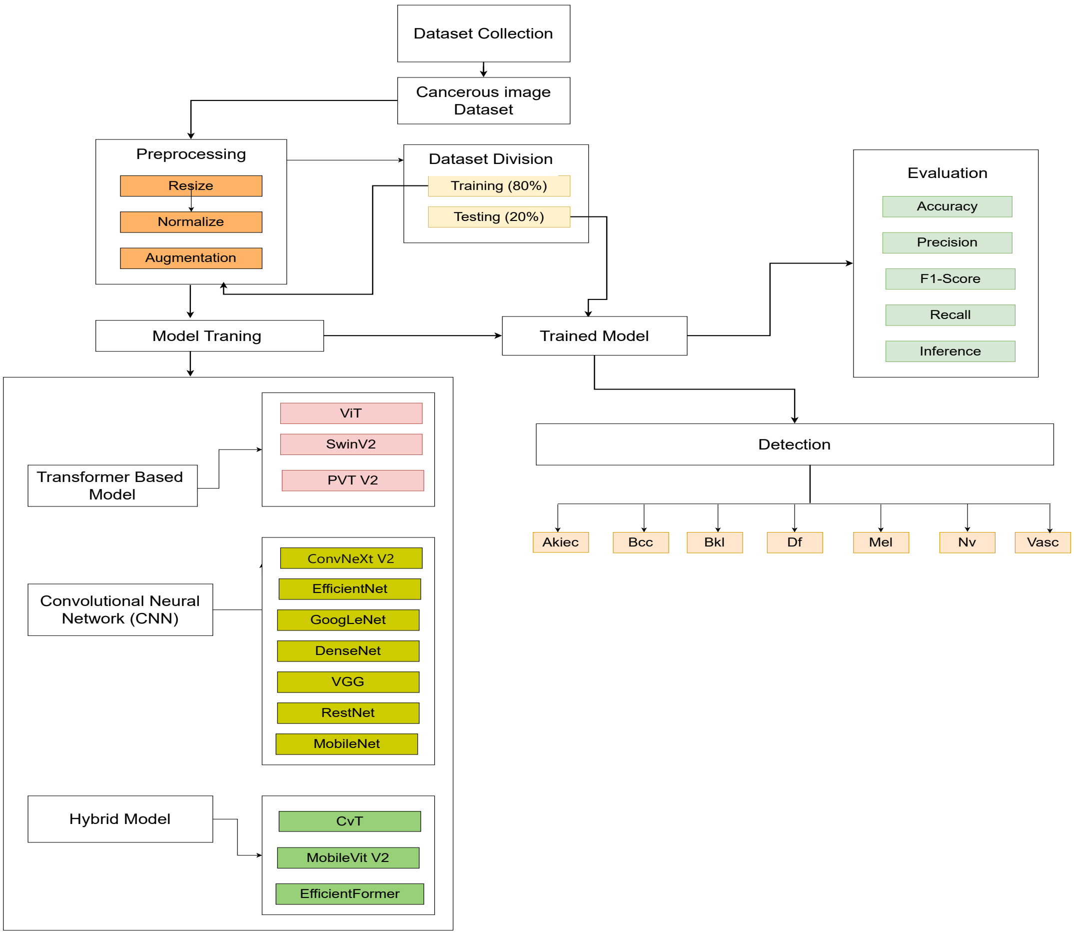

Fig. 1. Proposed method for skin cancer detection

carcinoma or Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc). Figure 2 shows the different kinds of lesions that are in the collection.

There are many things in the data set that make it easier to classify the diseases together. There are many different types of variables in the dataset, such as lesion ID, image ID, dx, dx type, age, sex, and localization. "dx" stands for the kind of disease that was spread [5].

B. Data pre-processing

Background information and noise arise from the image acquisition process and are already present in the query image. To remove the unwanted bits, preprocessing is necessary. The preprocessing stage usually removes unnecessary data. Noise causes pixels to display different intensity values than the image's real pixel values. For effective noise reduction, the initial preprocessing stage involves choosing the best strategy [22]. Data preprocessing includes image normalization, resizing, and augmentation for consistency and enhancement of model performance. Other methods, such as noise reduction and contrast enhancement, can boost image quality [23].

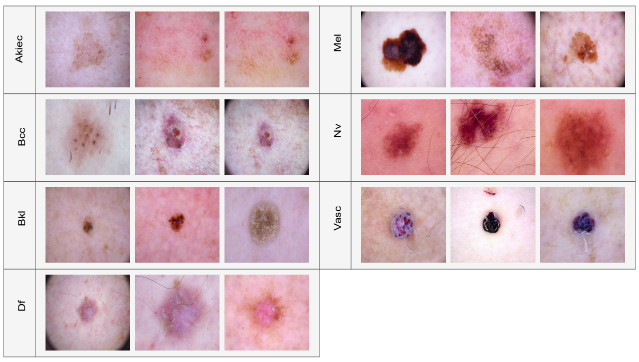

Fig. 2. Types of disease in dataset

1) Data Resize: The bilinear interpolation resampling filter method is employed to resize all images in the dataset to 224 × 224. It will significantly diminish the model's performance and expedite the refining process.

2) Normalization: During normalization, the picture pixels are scaled to a standard range, usually with a mean value of 0 and a standard deviation value of 1. To find the number from 0 to 1 for a pixel, divide it by 255. Therefore, the neural network's convergence is accelerated.

3) Dataset Partitioning: We divided the dataset into 80% training and 20% test sets.

C. Data Augmentation

Fig 3 shows that there is an imbalance in the number of different types of lesions in the data set. The data augmentation procedure increases the amount of data to be processed. This adjusts the data quantity for all classes about equally [24]. Data augmentation techniques, such as rotation (20°), width shift (0.2), height shift (0.2), zoom (0.2), horizontal and vertical flips, and fill mode ('nearest'), were used to mitigate the limited dataset size. The augmentation methods we applied to our training dataset are detailed in Table II. In the pre-processing phase, Data augmentation plays an important role in this work. It helped to equalize the data counts of each Lesion type.

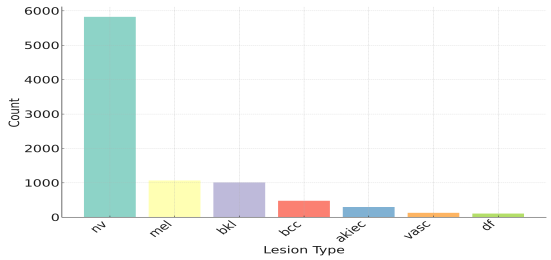

Fig. 3. Lesion type counts before augmentation

TABLE I
IMAGE AUGMENTATION TECHNIQUES

Transformations	Setting
Scale transformation	0.2
Rotation transformation	20°
Width shift transformation	0.2
Height shift transformation	0.2
Shear transformation	0.2
Zoom transformation	0.2
Horizontal flip	True
Vertical flip	True
Fill mode	'nearest'

Fig 4 shows the total number of lesions after augmentation.

D. Model Training and Classification

For the model training and classification part, transformer-based architecture, CNN-based network, and hybrid architecture are used. Table II shows the classification of models.

1) Transformers:

- **Vision Transformer :** Vision Transformers (ViT) have revolutionized computer vision by outperforming state-of-the-art models in several tasks. Using NLP approaches, this new architecture can attain or surpass typical CNN accuracy by translating picture patches into token sequences processed by transformer blocks. This breakthrough has shifted the paradigm in image processing [25].

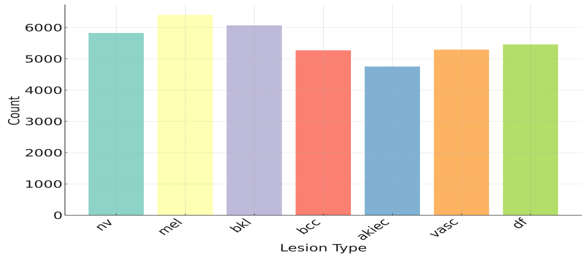

Fig. 4. Lesion type counts after augmentation

TABLE II
CLASSIFICATION OF MODELS INTO TRANSFORMERS, CNNS, AND HYBRID ARCHITECTURES

Category	Models
Transformers	ViT2, SwinV2, PVT V2, LLM
CNNs	ConvNeXt V2, VGG, GoogLeNet, DenseNet, ResNet, EfficientNet, MobileNet
Hybrid	CvT, EfficientFormer, MobileViTV2

- **Swinv2 :** In response to concerns with training stability and resolution gaps between pre-training and fine-tuning, the SwinV2 Transformer was created. The original Swin Transformer had training instability because of high activation output amplitude disparities across network layers [26].
- **PVT V2** PVT v2's pyramid structure captures multi-scale characteristics better than other visual Transformers, making it better for dense pixel prediction like polyps. PVT v2 further captures local continuity information of polyp pictures via overlapping patch embedding, which helps mine polyp borders and irrelevant backgrounds [27].

2) CNNs:

- **ConvNeXt V2 :** Swin Transformer network structure and ResNet-50 architecture were used to create the ConvNeXt model. ConvNeXt outperforms Swin Transformer on COCO detection and ADE20K. The ConvNeXt model's creators trained ResNet-50 using ViTs' technique, which improved outcomes. From this baseline, tests were done. ConvNeXt V2 is a new model that combines self-supervised learning and architectural enhancements [28] [29].
- **VGG** : The primary uses of VGG are in identification and detection. The amount of layers is what makes anything "deep." Innovative object identification models are built on top of the VGG architecture. Furthermore, it remains one of the most widely used image recognition systems to this day [30].
- **GoogLeNet:** A convolutional neural network (CNN) with 22 layers is called GoogLeNet. GoogLeNet has

nine linearly fitted inception modules. The global average pooling architecture replaces fully linked layers with map feature averages. Currently, GoogLeNet is used for computer vision tasks such as face detection, adversarial training, and more [31].

- **DenseNet** DenseNet uses dense layer connections, minimizes parameters, improves propagation, and reuses features. Improved parameter efficiency enables quicker and simpler network training [32]. Additionally, neural network model characteristics maintain their distribution. DenseNet improves ResNet's residual application by linking all tightly linked (little to no redundancy) prior layers to the layer in front of them [33].
- **ResNet** ResNet refers to a residual network, an ANN type. ResNet trains deep neural networks with over 500 layers and produces impressive performance. The amount of parameters used depends on how many layers there are in the network. It relies on batch normalization. To improve network performance, batch normalization is used to tune the input layer [34].
- **EfficientNet** CNN network EfficientNet is known for its speed and parameter efficiency. EfficientNet design aims to improve classification accuracy with fewer parameters. EfficientNet architecture, built in 2020, obtained an accuracy of 0.844 in classifying the ImageNet dataset. EfficientNet improves efficiency by scaling depth, breadth, and resolution while reducing architectural size [35]. The EfcientNet family consisted of seven CNN models, names EfcientNet-B0 to EfcientNet-B7 [36].
- **MobileNet** There are 28 levels in the MobileNet network. The normal MobileNet design has about 4.2 million parameters. The number of parameters can be cut even more by changing the width multiplier hyperparameter. The skin lesion image is first pre-processed to get rid of any copies, and then it is sent to MobileNet to be classified [37]

E. Hybrid Architecture :

- **Cvt** : The Convolutional Vision Transformer (CvT) design makes use of convolutions. The CvT design incorporates convolution to improve the model's sensitivity to fine-grained pictures and to effectively collect local information, which are two crucial components of the ViT architecture [38].
- **EfficientFormer:** EfficientFormer utilizes 4D MetaBlocks from PoolFormer to efficiently learn local representations and 3D MetaBlocks from self-attention to encode global context. Two design considerations restrict EfficientFormer's performance. The system utilizes inefficient token mixing and only uses 3D MetaBlocks in the last stage owing to the quadratic complexity of multi-headed self-attention (MHSA) [39].
- **MobileViT :** MobileViT combines CNN and ViT capabilities. MobileViT perceives transformers as convolutions, thus it can use inductive biases and long-range dependencies to construct a lightweight mobile network. Although MobileViT networks have fewer parameters and outperform light-weight CNNs [40].

IV. Result and Discussion

This section discusses how well different models work for classifying skin lesions. The HAM10000 dataset was used for the research. Every model's results are shown in Table III so they can be easily compared. Performance factors like the F1-score, precision, accuracy, and recall are used to compare different designs. From the confusion matrix, we can find four expected values for the results. We used those numbers (TN, TP, FN, and FP) to figure out the algorithms' Accuracy, Recall, Precision, F1-score, and Error rate. Inference time is also calculated with the hardware setup which has a direct impact on calculating inference time. For this research work, we have tuned the hyperparameter: learning rate 1e-05, to improve the performance and accuracy of the models.

TABLE III
Analysis of Models on Performance Metrics

Category	Model	Accuracy	Precision	Recall	F1-score
Transformer	SwinV2	92.74%	0.85	0.75	0.79
	ViT2	92.66%	0.81	0.74	0.77
	PVT V2	91.75%	0.77	0.74	0.75
CNN	ConvNeXt V2	93.2%	0.85	0.74	0.79
	VGG	91.56%	0.76	0.74	0.75
	GoogLeNet	90.11%	0.77	0.69	0.72
	DenseNet	91.84%	0.78	0.74	0.75
	ResNet	91.65%	0.80	0.70	0.73
	EfficientNet	91.65%	0.79	0.78	0.78
	MobileNet	91.11%	0.79	0.74	0.76
Hybrid	CvT	92.20%	0.75	0.76	0.74
	EfficientFormer	91.56%	0.76	0.74	0.75
	MobileViTV2	90.75%	0.77	0.70	0.72

Measuring how frequently a machine learning model predicts the result properly is called accuracy. In this research work, ConvNeXt V2 gives the highest accuracy, 93.2% for the HAM10000 Dataset which is depicted in Fig 5. ViT2, Cvt, and Swinv2 perform similar in terms of accuracy which is 92%.

The statistical tests are measured by the F1-score. F1-score uses Precision and Recall to calculate the prediction accuracy. The F-score may also be calculated using the weighted average of recall and accuracy. ConvNeXt V2 and Swinv2 have the highest F1-score of 0.79. ViT2, which has an F1-score of 0.77, comes in second. Besides, DenseNet, PVT V2, EfficientFormer, and VG generate an almost similar F1-score, which is 0.75.

The recall is obtained by dividing the number of accurate guesses by the total number of predictions. Recall evaluates the completeness of positive predictions. EfficientNet-B0 possesses the highest recall value of 0.78, demonstrating that, out of all actual positives, it has the best capability to recognize true positives. SwinV2 has the second-highest value of 0.75. ViT2, PVT V2, EffientFormer, VGG, and MobileNet follow with a recall value of 0.74. The second one in terms of precision is ViT2, a score of 0.81.

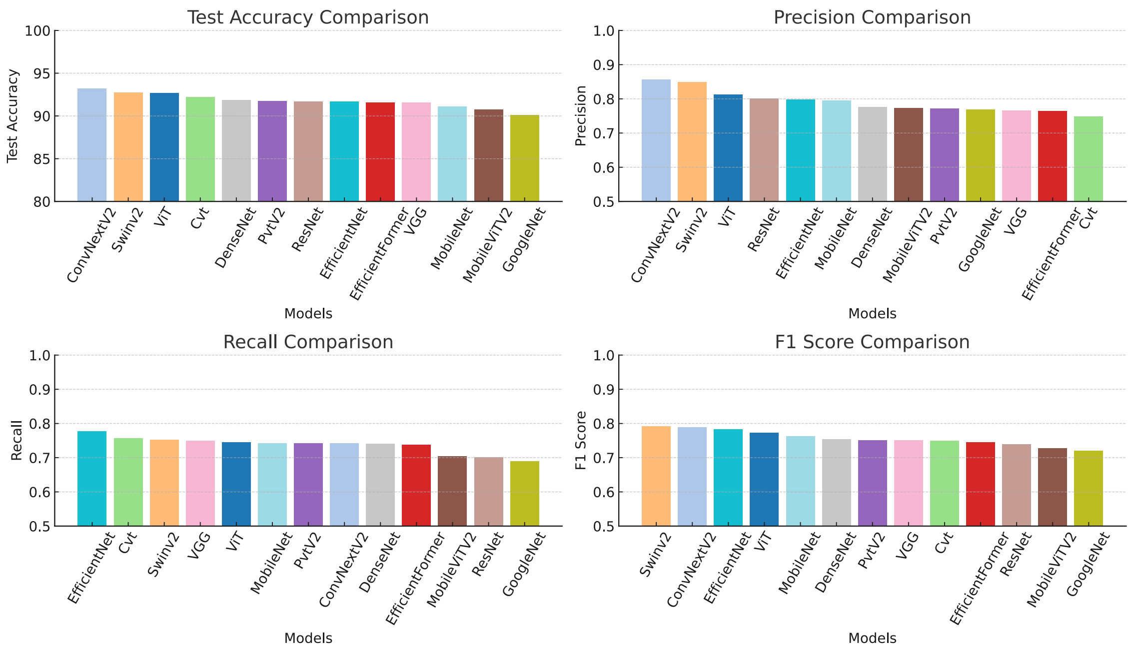

Fig. 5. Comparison Scores of the Implemented Models

TABLE IV
CATEGORY-WISE COMPARISON OF MODELS WITH EXISTING WORK

Category	Source	Model Name	Accuracy
Transformer	Arshed et al. [7]	Vision Transformer	92.14%
	Zhenhao Zhao [2]	ViT	82.64%
	Zhenhao Zhao [2]	DeepViT	84.20%
	Proposed System	SwinV2	92.74%
CNN	Zhenhao Zhao [2]	VGGNet	92.61%
	Zhenhao Zhao [2]	ResNet	93.26%
	Ali et al. [10]	EfficientNet B4	87.90%
	Lariba and Akanzawon [11]	ResNet50	86.69%
	Lariba and Akanzawon [11]	DenseNet201	86.91%
	Fraiwan and Faouri [15]	GoogLeNet	73.4%
	Fraiwan and Faouri [15]	DenseNet201	82.9%
	Fraiwan and Faouri [15]	MobileNetV2	74.9%
	Fraiwan and Faouri [15]	EfficientNetB0	76.7%
	GURURAJ et al. [19]	DenseNet169	91.2%
	GURURAJ et al. [19]	ResNet50	83.0%
	Proposed System	ConvNeXt V2	93.2%
Hybrid	Shehzad et al. [16]	EfficientNetV2S + Swin-Transformer	99.10%
	Proposed System	CvT	92.20%

The precision metric is calculated by dividing the total number of correctly predicted forecasts by the total number of returned predictions. The accuracy of positive predictions is measured. ConvNeXt V2 and SwinV2 have a precision value of 0.85, highlighting that, among all the estimates, it has the highest percentage of true positives. The EfficientNet-B0 is the third-best model with a precision value of 0.79.

Inference time in deep learning is the amount of time it takes for a learned model to make predictions based on new data that it hasn't seen before. The inference time of a deep learning

TABLE V
HARDWARE CONFIGURATION

Specification	Values
Architecture	x86_64
CPU operation model	32 and 64 bit
No of CPU	64
Processor	Intel Xeon Silver 4216 CPU
Threads and COre	2,2
GPU model	Tesla v100SXM2
No of GPU	1
Memory Usage	3,712 MiB, 32,510 MiB

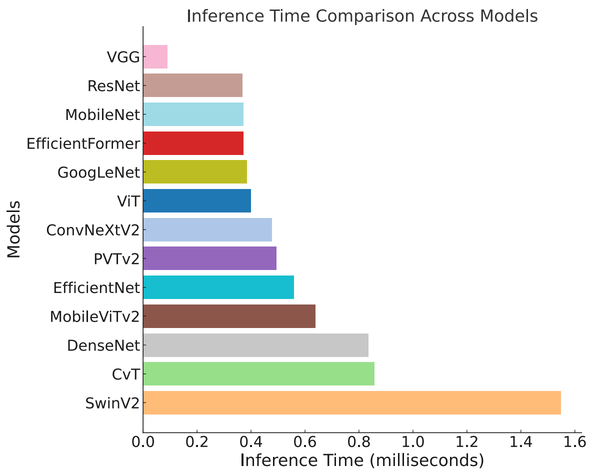

Fig. 6. Average Inference Time

model is influenced by various factors, including the model architecture's complexity, the size of the input data, and the hardware on which the model is running (e.g., CPU, GPU, or specialized hardware like TPUs). In this work, we calculated the average inference time of each mode which is illustrated in Fig 6. Among all models, VGG took the lowest time to predict, which is 0.00009 sec. SwinV2 predicts new and unseen data on average in 0.000154 sec, which is the longest time among all algorithms. Cvt and DenseNet have the second highest average inference time of 0.00088 sec, 0.00084 sec. For our research work, The hardware configuration includes an x86_64 architecture with both 32-bit and 64-bit CPU operation models. The system has 64 CPUs, specifically Intel Xeon Silver 4216 processors, each with 2 threads and 2 cores. For GPU, the system uses a single Tesla V100-SXM2 model. Memory usage is recorded at 3,712 MiB out of 32,510 MiB. The hardware configuration is depicted in Table V.

Moreover, in this work, we have analyzed several works that have used the HAM10000 dataset and employed the algorithms we used in this system. Table IV shows that SwinV2 gives the highest accuracy in this study among the transformer models compared to the other research papers analyzed here. In this work, several CNN models were trained. Based on the existing research work, the RestNet model gave an accuracy of 93.26% [2], which is a significant accuracy. However, ConvNeXt V2 gave the highest accuracy, 93.2%, in our proposed system. Furthermore, from Table IV, a research work that utilized EfficientNetV2S and SwinTransformer, where they got an accuracy of 99.10% [16], whereas, in our work, CvT has the highest accuracy among the hybrid models.

V. Conclusion and Future Work

This work looked into several deep learning algorithms categorized into three genres: Transformer, CNN, and Hybrid. These models detect skin cancer in seven categories: Df, Bkl, Bcc, Vasc, Mel, and Nv. To feed the deep learning models, datasets, such as data resizing, normalization, partitioning, and augmentation, are pre-processed to improve their quality. This study aims to provide a guide for diseased data, which can be used to invent integrated algorithms by employing these trendy algorithms implemented here to classify lesion types. However, the HAM10000 dataset has several limitations. Expanding the size of the HAM10000 dataset would yield numerous benefits for this research endeavor. Utilizing a large dataset enhances the precision and enables the models to generalize more effectively. Furthermore, the number of classes in this dataset is highly imbalanced which poses a great drawback to using this for deep learning models.

Furthermore, the models are evaluated on different evaluation indicators: confusion matrix, accuracy, precision, recall, and inference time. After calculating the value, it is clear that the ConvNeXt V2 model has the highest accuracy, 93.2%, on the HAM10000 dataset. Moreover, the inference time is the lowest in the VGG model. However, ConvNeXt V2, Swinv2, DenseNet, PVT V2, and EfficientFormer also showed good F1-score, precision, and recall values. The research is helpful for knowledge of three different categories of models for image classification, especially for further research on the detection of diseases. Additionally, In the future, we will try to gather real-time data on diseases along with modified algorithms by combining several deep learning models, which will be a key goal for future research and make a great impact in the treatment of skin cancer.

References

[1] K. Sambyal, S. Gupta, and V. Gupta, "Skin cancer detection using resnet," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, 2022.

[2] A. Imran, A. Nasir, M. Bilal, G. Sun, A. Alzahrani, and A. Almuhaimeed, "Skin cancer detection using combined decision of deep learners," *IEEE Access*, vol. 10, pp. 118198–118212, 2022.

[3] E. H. Mohamed, A. F. Abubakr, N. Abdu, M. Khalil, H. Kamal, M. Youssef, H. Mohamed, and M. ElSayed, "A hybrid deep learning framework for skin cancer classification using dermoscopy images and metadata," 2023.

[4] N. Rezaoana, M. S. Hossain, and K. Andersson, "Detection and classification of skin cancer by using a parallel cnn model," in *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pp. 380–386, IEEE, 2020.

[5] D. Moturi, R. K. Surapaneni, and V. S. G. Avanigadda, "Developing an efficient method for melanoma detection using cnn techniques," *Journal of the Egyptian National Cancer Institute*, vol. 36, no. 1, p. 6, 2024.

[6] S. N. Hayat, R. Indraswari, *et al.*, "Skin cancer detection approach using convolutional neural network artificial intelligence," *International Journal of Informatics and Information Systems*, vol. 7, no. 2, pp. 46–54, 2024.

[7] M. A. Arshed, S. Mumtaz, M. Ibrahim, S. Ahmed, M. Tahir, and M. Shafi, "Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models," *Information*, vol. 14, no. 7, p. 415, 2023.

[8] M. S. Islam, S. Nag, A. Dutta, M. Ahmed, F. F. Niloy, and A. K. Roy-Chowdhury, "Active learning guided federated online adaptation: Applications in medical image segmentation," 2023.

[9] G. Yang, S. Luo, and P. Greer, "A novel vision transformer model for skin cancer classification," *Neural Processing Letters*, vol. 55, no. 7, pp. 9335–9351, 2023.

[10] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using efficientnets–a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.

[11] C. Agyenta and M. Akanzawon, "Skin lesion classification based on convolutional neural network," *Journal of Applied Science and Technology Trends*, vol. 3, no. 01, pp. 21–26, 2022.

[12] A. Z. Yutra, J. Zheng, X. Li, and A. Endris, "Skinaacn: An efficient skin lesion classification based on attention augmented convnext with hybrid loss function," in *Proceedings of the 2023 7th International Conference on Computer Science and Artificial Intelligence*, pp. 295–300, 2023.

[13] Z. Li, Y. Han, and X. Yang, "Multi-fundus diseases classification using retinal optical coherence tomography images with swin transformer v2," *Journal of Imaging*, vol. 9, no. 10, p. 203, 2023.

[14] L. Zheng, J. Fang, X. Tang, H. Li, J. Fan, T. Wang, R. Zhou, and Z. Yan, "Pvt-cov19d: Pyramid vision transformer for covid-19 diagnosis," *arXiv preprint arXiv:2206.15069*, 2022.

[15] M. Fraiwan and E. Faouri, "On the automatic detection and classification of skin cancer using deep transfer learning," *Sensors*, vol. 22, no. 13, p. 4963, 2022.

[16] K. Shehzad, T. Zhenhua, S. Shoukat, A. Saeed, I. Ahmad, S. Sarwar Bhatti, and S. A. Chelloug, "A deep-ensemble-learning-based approach for skin cancer diagnosis," *Electronics*, vol. 12, no. 6, p. 1342, 2023.

[17] M. S. I. Sajol and A. S. M. J. Hasan, "Benchmarking cnn and cutting-edge transformer models for brain tumor classification through transfer learning," in *2024 IEEE 12th International Conference on Intelligent Systems (IS)*. Accepted for publication. TechRxiv. July 23, 2024.

[18] S. Pan, Z. Tian, Y. Lei, T. Wang, J. Zhou, M. McDonald, J. D. Bradley, T. Liu, and X. Yang, "Cvt-vnet: a convolutional-transformer model for head and neck multi-organ segmentation," in *Medical imaging 2022: computer-aided diagnosis*, vol. 12033, pp. 928–935, SPIE, 2022.

[19] H. L. Gururaj, N. Manju, A. Nagarjun, V. M. Aradhya, and F. Flammini, "Deepskin: a deep learning approach for skin cancer classification," *IEEE Access*, vol. 11, pp. 50205–50214, 2023.

[20] Y. Ding, Z. Yi, M. Li, J. Long, S. Lei, Y. Guo, P. Fan, C. Zuo, and Y. Wang, "Hi-mvit: A lightweight model for explainable skin disease classification based on modified mobilevit," *Digital Health*, vol. 9, p. 20552076231207197, 2023.

[21] A. Obaid, A. Shawkat, and N. Salih, "A powerful deep learning method for skin cancer detection," *Journal of Autonomous Intelligence*, vol. 7, 11 2023.

[22] V. D. Midasala, B. Prabhakar, J. K. Chaitanya, K. Sirnivas, D. Eshwar, and P. M. Kumar, "Mfeuslnet: Skin cancer detection and classification using integrated ai with multilevel feature extraction-based unsupervised learning," *Engineering Science and Technology, an International Journal*, vol. 51, p. 101632, 2024.

[23] F. Olaoye, "Investigating the application of machine learning and deep learning for skin cancer detection," 2024.

[24] S. Kanrar and H. Chhabra, "Skin cancer detection using convolutional neural networks," in *Proceedings of International Conference on Advanced Computing Applications*, (Singapore), pp. 469–482, Springer Singapore, 2022.

[25] F. Chen, Z. Luo, L. Zhou, X. Pan, and Y. Jiang, "Comprehensive survey of model compression and speed up for vision transformers," *arXiv preprint arXiv:2404.10407*, 2024.

[26] K. Fitzgerald, J. Bernal, A. Histace, and B. J. Matuszewski, "Polyp segmentation with the fcb-swinv2 transformer," *IEEE Access*, 2024.

[27] Z. Li, L. Zhang, S. Yin, and G. Zhang, "Mscff-net : Multi-scale context feature fusion network for polyp segmentation," 2024.

[28] Y. Xu, J. Li, L. Zhang, H. Liu, and F. Zhang, "Cntcb-yolov7: An effective forest fire detection model based on convnextv2 and cbam," *Fire*, vol. 7, no. 2, p. 54, 2024.

[29] M. S. I. Sajol, A. S. M. J. Hasan, M. S. Islam, and M. S. Rahman, "A convnext v2 approach to document image analysis: Enhancing high-accuracy classification," in *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*, (Louisiana, USA), IEEE, 2024. Accepted.

[30] H. Tabrizchi, S. Parvizpour, and J. Razmara, "An improved vgg model for skin cancer detection," *Neural Processing Letters*, vol. 55, no. 4, pp. 3715–3732, 2023.

[31] S. Barman, M. R. Biswas, S. Marjan, N. Nahar, M. S. Hossain, and K. Andersson, "Transfer learning based skin cancer classification using googlenet," in *Machine Intelligence and Emerging Technologies* (M. S. Satu, M. A. Moni, M. S. Kaiser, and M. S. Arefin, eds.), (Cham), pp. 238–252, Springer Nature Switzerland, 2023.

[32] M. A. Wakili, H. A. Shehu, M. H. Sharif, M. H. U. Sharif, A. Umar, H. Kusetogullari, I. F. Ince, and S. Uyaver, "Classification of breast cancer histopathological images using densenet and transfer learning," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 8904768, 2022.

[33] S. Bin, J. Qingqing, and S. Qingbing, "Image dehazing algorithm based on fc-densenet and wgan," *Journal of Frontiers of Computer Science & Technology*, vol. 14, no. 8, p. 1380, 2020.

[34] S. Sasikala, S. Shivappriya, *et al.*, "Towards improving skin cancer detection using transfer learning," *Biosci Biotechnol Res Commun*, vol. 13, no. 11, pp. 55–60, 2020.

[35] M. UÇan, B. Kaya, and M. Kaya, "Multi-class gastrointestinal images classification using efficientnet-b0 cnn model," in *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 1–5, IEEE, 2022.

[36] X. Chen, X. Pu, Z. Chen, L. Li, K.-N. Zhao, H. Liu, and H. Zhu, "Application of efficientnet-b0 and gru-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions," *Cancer Medicine*, vol. 12, no. 7, pp. 8690–8699, 2023.

[37] V. A. O. Nancy, P. Prabhavathy, M. S. Arya, and B. S. Ahamed, "Comparative study and analysis on skin cancer detection using machine learning and deep learning algorithms," *Multimedia Tools and Applications*, vol. 82, no. 29, pp. 45913–45957, 2023.

[38] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22–31, 2021.

[39] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17425–17436, October 2023.

[40] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," 2022.