

Explainable AI (XAI) driven Skin Cancer detection using Transformer and CNN based architecture

by

Faiza Radiah

19101288

Kabasum Rahman

19101645

Lasania Asadullah

19101144

Md. Sohanur Rahman Sohan

19301229

Jaki Ahmed

19301161

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
Fall 2023

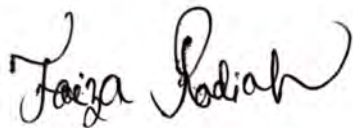
© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

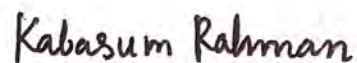
1. The thesis submitted is our original work while completing the degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Faiza Radiah

19101288




Kabasum Rahman

19101645



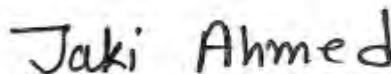
Lasania Asadullah

19101144



Md. Sohanur Rahman Sohan

19301229



Jaki Ahmed

19301161

Approval

The thesis titled “Explainable AI (XAI) driven Skin Cancer detection using Transformer and CNN based architecture” submitted by:

1. Faiza Radiah (ID: 19101288)
2. Kabasum Rahman (ID: 19101645)
3. Lasania Asadullah (ID: 19101144)
4. Md. Sohanur Rahman Sohan (ID: 19301229)
5. Jaki Ahmed (ID: 19301161)

Of Fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on Jan 22, 2024.

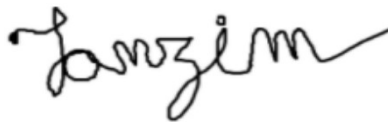
Examining Committee:

Supervisor:



Md. Ashraful Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:



Md. Tanzim Reza
Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Co-Ordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Skin Cancer is a cancer form that has become very prevalent in recent times and, if left untreated, has the potential to cause premature death. That is why early diagnosis and treatment are important to cure this disease. For this, we can use Machine Learning based methods to effectively impact the identification and categorization of skin cancer. Previously it was seen that the CNN models had a notable impact on the performance of the classification tasks. However, Vision transformers (ViT) are also the solution chosen by the researchers which have displayed significant performance in classification works. To make the outcomes of diverse data as distinct as feasible, contrastive learning is utilized to make similar skin cancer data for encoding similarly. The categorization of skin cancer depending upon multimodal data is made possible by the transformer network's exceptional performance in natural language processing and field of vision. In this paper, we have offered a detailed analysis of VGG-16, a CNN architecture, and ViT, a transformer-based method to classify skin lesion images for aiding the early diagnosis of skin cancer. The findings indicate that the VGG-16 model attained an accuracy of 82.14%, whereas the Vision Transformer achieved a slightly lower accuracy of 76.15%. A modified version of the original vision transformer, the shifted patch tokenization, and locality self-attention modified Vision transformer showed an accuracy of 74.55% with expectations for further improvement in the future. Moreover, nowadays people have to choose a model from several other models to solve an issue, and as the model keeps on improving, it becomes very difficult to understand how the model works internally. So, for this reason, Explainable Artificial Intelligence (XAI) is introduced to give an idea of a human-readable explanation for the decision-making process of a model. This will certainly benefit cosmetologists, health researchers, research scientists, and researchers working in various areas and offer patients more convenience.

Keywords: Skin cancer, Deep Learning, CNN, VGG-16, ViT, XAI, Dermatoscopy, Augmentation, GradCam

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption.

Secondly, to our supervisor Dr. Md. Ashraful Alam sir and our co-supervisor Md Tanzim Reza for their kind support and advice in our work. They helped us whenever we needed help.

And finally to our parents without their support, it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	iv
Dedication	v
Acknowledgment	v
Table of Contents	vi
List of Figures	viii
List of Tables	1
1 Introduction	2
1.1 Background of Skin Cancer	2
1.2 Problem Statement	3
1.3 Research Objectives	4
1.4 Research Orientation	4
2 Literature Reviews	5
3 Research methodology	9
3.1 Dataset	10
3.1.1 Dataset Source	10
3.1.2 Description of the Dataset	10
3.2 Data Preprocessing	13
3.2.1 Data Resize	13
3.2.2 Data Augmentation	13
3.3 Data Splitting	13
3.3.1 Training Set	13
3.3.2 Validation Set	13
3.3.3 Testing Set	13
3.4 Used Architectures	14

4	Implementation And Result Analysis	17
4.1	Experimental Setup	17
4.1.1	Confusion Matrix	17
4.1.2	Accuracy	18
4.1.3	Precision	18
4.1.4	Recall	18
4.1.5	F1 score	18
4.2	Performance Measure	19
4.3	Result Analysis	21
5	GradCam	23
6	Conclusion	25
	Bibliography	28

List of Figures

3.1	Our proposed methodology	9
3.2	Sample of ISIC 2019 dataset: A) scc, B) bkl, C) akiec, D) vasc, E) bcc, F) mel, G) nv, and H) df	11
3.3	Internal Architecture of VGG-16	14
3.4	Internal Architecture of Vision Transformer	15
3.5	Internal Architecture of Shifted Patch Tokenization and Locality Self Attention modified ViT model (ViT-STP)	16
4.1	Loss and accuracy of VGG-16 model	19
4.2	Loss and accuracy of ViT model	19
4.3	Loss and accuracy of vanilla ViT model	19
4.4	Loss and accuracy of Shifted Patch Tokenization and Locality Self Attention modified ViT model	20
4.5	Classification report and confusion matrix of VGG-16 model	20
4.6	Classification report and confusion matrix of ViT model	20
4.7	Classification report and confusion matrix of vanilla ViT model	21
4.8	Classification report and confusion matrix of Shifted Patch Tokenization and Locality Self Attention modified ViT model	21
5.1	GradCam outputs for 4 different classes (BCC, BKL, MEL, NV) on VGG-16 model	24

List of Tables

3.1	Class-wise data distribution	10
4.1	Confusion Matrix	17
4.2	Result Analysis	22

Chapter 1

Introduction

1.1 Background of Skin Cancer

Skin cancer primarily is the uncontrolled enlargement of epithelial cells that predominantly affects the skin that has been exposed to UV rays. Skin cancer comes in three main varieties: Basal Cell Carcinoma, Squamous Cell Carcinoma, and Melanoma. However, skin cancer of the typical form might develop across areas of the body which is not commonly revealed in direct sunlight. It is also seen that skin cancer may affect people of any type of skin tone, but it is most common among individuals with a darker complexion. It most likely appears on the parts of the body that are in direct sunlight for example face, head, scalp, mouth, ears, nose, throat, chest, forearm, and legs. However, the skin right below our fingernails or toenails, hands, or pubic regions are examples of sites where it can also appear. Furthermore, those with darker skin tones tend to be more susceptible to melanoma on the palms of their hands and bottom of their feet, which are typically not exposed to the sun.

By keeping an eye out for unusual changes on the surface of the skin, one may discover the early indications of skin cancer. Some types of skin cancer and their symptoms are as follows:

1. Basal Cell Carcinoma typically develops in the outer skin that usually has direct contact with the sun, such as the face or neck area. Pearly or waxy lumps, plain, skin-colored, or fleshy type sores that are brown, and hemorrhagic or blistering sores that recover and recur are signs of the condition.
2. Those having tanned complexions are often more prone to develop Squamous Cell Carcinoma than those who are not regularly under the sun. It develops in sun-exposed parts of the body. Firm, red nodules and flat sores with a crusty and rough surface are signs of the condition.
3. Any skin tone can be impacted by melanoma. The symptoms include huge brown patches with black freckles and moles which fluctuate in size, and color, and little sores with an erratic borderline that are red, pink, white, blue, or blue-black.

If skin cancer is found in its earliest stages, we have the highest chance of effectively treating it. Mainly, for skin cancer, the major possible causes are lighter skin tone,

extensive UV contact, sunburn, moles, climatic factors, family history, weak immunity system, and exposure to certain substances like arsenic.

Again, it is evident that the rapidly aging population would be blamed for the considerable rise in skin cancer cases over the past several years. With this growing number of skin cancer incidences, dermatologists and medical institutions face greater challenges in diagnosing it accurately. Dermatoscopy has become very popular recently among medical experts for detecting skin lesions but it is quite expensive and requires highly skilled and experienced professionals to make the correct diagnosis. To meet the challenges of dermatoscopy, AI has advanced remarkably in the realm of medical imaging. On the other hand, a Deep Learning algorithm called CNN (Convolutional Neural Network) is frequently employed in image categorization and has demonstrated effective performance in diagnosis. However, CNN lacks in learning long-range spatial relations sometimes and often does not focus on the important part of the images.

Therefore, we aim to implement a popular CNN-based architecture, VGG-16, and two transformer-based models named Vision Transformer and Shifted Patch Tokenization and Locality Self-attention modified Vision Transformer and then provide some visual explanation of how our model works through Explainable AI on the available dataset for classifying skin lesion images to diagnose skin cancer as timely and as accurately as possible.

1.2 Problem Statement

Among all the organs in the human body, the skin is the largest one which covers the whole body and protects it from external harm. Because the skin is our body's external defense system, it is prone to a variety of conditions. According to a study, some form of skin condition affects about one-third of people worldwide. Despite its high occurrence rate, people often pay less attention to skin-related issues considering them minor infections. Patients feel uncomfortable to share about their skin conditions with the doctor and tend to hide those as they can occur in any part of the body. Moreover, skin diseases are still considered taboo and there are many misconceptions associated with them. People avoid skin disease patients thinking that they will spread their disease which makes the patient reluctant to go to the doctor and treat their issues. However, some of the conditions can worsen over time and become cancerous if not treated timely.

Therefore, diagnosing skin cancer at the earliest stage is crucial to reduce the patient's lifetime risk. It is estimated that 99% of patients with melanoma will survive for five years if detected early but the percentage decreases significantly with the delay in detection. It is to be mentioned that to train data-driven models properly, a large amount of data is needed to achieve their full potential. However, there is always a limitation in the availability of skin-related data. That is because skin abnormalities can appear in any part of the body and the data becomes more confidential than any other medical data. Access to skin-related data is strictly regulated and medical centers don't share it with others because of privacy issues. Thus working with skin-related data becomes more challenging. Although there are traditional

methods for detecting skin lesions like dermatoscopy, they are time-consuming and might produce inaccurate results, leading to treatment delays. So, developing a consistent way for early skin lesion detection has become very essential. For this, we plan to train a CNN model, VGG-16, and two transformer models named vision transformers on our dataset and assess which model achieves the best results to classify skin lesions for diagnosing cancer properly with the least execution time.

Again, the use of the desired methods for addressing AI's "black box" dilemma has increased recently. It is a result of the deep learning networks' intricate architecture, which includes numerous hidden layers and makes the internal logic of the network difficult to understand. Although the model's architecture can be easily visualized at a detailed level, that does not provide us with a complete view of what the model perceives when it solves a specific issue. So it arises the concept of Explainable AI (XAI) which is the idea of providing a human-readable explanation for a model's decision-making process. These days, people can choose from hundreds of different intricate model architectures to solve only one issue, and as model performance keeps on improving, it becomes more difficult to understand how the models work internally. Comprehending and visualizing the internal processes of models, especially those pivotal in making classification decisions, is essential, particularly when human lives are at stake. For this reason, Explainable AI is used for exploring and understanding what a model sees while deciding on a classification task.

1.3 Research Objectives

The objective of the research would be to make a comparison analysis between CNN and transformer-based architectures on dermoscopic images for advanced skin cancer diagnosis. To enhance the model's performance, the data will be first processed by different pre-processing techniques such as image resizing and augmentation. We intend to apply a predefined deep learning model like VGG-16 and also two of the vision transformer models on our dataset and further use XAI, specifically, Grad-Cam on the best-performing model to visualize its classification decision. Our objectives for the research are:

1. Implementation of early detection of skin cancer using CNN and vision transformer architectures.
2. Implementation of XAI to comprehend model performance, assist the cognitive comprehension of the model's functionality, and provide visual information regarding the model's emphasis.

1.4 Research Orientation

Following this chapter, we have presented the literature review in Chapter 2 and proposed methodology, dataset description, and also the description of the models along with some visual representations in Chapter 3. In Chapter 4, we have presented the implementation and result analysis. Later, in chapter 5, we illustrated a short description of GradCam and lastly, we concluded in chapter 6.

Chapter 2

Literature Reviews

(Cai et al.,2022) [1] used two multimodal datasets of images depicting skin disorders and metadata for clinical trials. The model they initiated was made up of a decoder and 2 encoders inspired by the transformer framework. They compared multiple ViTs against some widely used CNN architectures where NesT achieved higher accuracy (75%) than DenseNet121 and ResNet101 (66%) within the private dataset hence, during the study of the private dataset, it acted as the network's support structure. For the metadata, SLE and MA block proved to be more suitable than the one hot encoder method in this case. When the proposed model was put up against the other that takes input images, the accuracy reached 0.816 from 0.75. It is to be noted that in the medical sector, image processing and segmentation play an important role. To detect Melanoma, a type of skin cancer, the melanoma parameters can be used for image segmentation and feature stages. The picture is identified either as cancer-free or as a melanoma cancer lesion using the texture, size, and shape retrieved from feature parameters. According to the authors [2], computer vision can be useful for image diagnosis in healthcare. It is necessary to create computed diagnostic techniques to assist individuals in the early diagnosis of melanoma to cope with this issue. Segmenting skin lesions is the initial step. Feature and pattern analysis processes must be executed as the next crucial step to diagnose the damaged region. Many methods for skin lesion segmentation have been created by integrating various CNN architectures with multi-scale data. Such approaches either require extra labeling or rely on large, practically useless parameters. Therefore, in [3] researchers offered a hybrid form of segmentation algorithms which is the combination of ViT and ConvNet that showcased high accuracy in both the localization of infection and the identification of skin lesions.

According to [4], the researchers believe that to prevent the spread of fatal melanoma skin cancer, detection of the skin lesions at the earliest possible time is crucial and dermatoscopy is the best way for that. However, automatic skin lesion segmentation becomes difficult when the lesions are similar in color or visual pattern. To exceed the limitations of CNN on skin lesion segmentation, the researchers have applied convolution-deconvolution-based (U-NeT, V-NeT) and attention-based (Attention U-NeT, TransUNeT, Swin-UNeT) models on the ISIC 2018 dataset in a modified way for each specific task. While comparing with the separate U-NeT-based methods, attention-based methods integrated with U-NeT properties achieved better results in distinguishing pigment regions. In the paper [5], researchers applied SVM, KNN,

and CNN models to test the dataset and among the three models, CNN obtained the best results of 85.5%. On a histopathologic cancer diagnostic dataset in [6], the researchers evaluated the effectiveness of 14 pre-trained ImageNet models, each of which is classified as a fine-tuned model, naive model, or feature extractor model. The study showed that Resnet101 had a high recall rate whereas Densenet161 displayed more precise performance. Densenet161 had an AUC score of 0.9924 and an f1 Score of 0.95, which outperformed the other designs within the feature extractor model. In the paper [7], a GWO technique-based CNN was created to accurately identify the skin cancer type based on input photos. By using a suitable encoding technique, the approach optimized the CNN hyperparameters using the Grey Wolf Optimization algorithm. The researchers combined all of the available skin-level characteristics and provided a thorough forecast of the patient’s health using a variety of deep-learning neural networks and data analysis approaches [8]. To identify melanoma, again CNNs that were pre-trained on pictures of skin lesions were selected and fine-tuned. The article [9] presented a more complete approach to dermoscopy image-based skin cancer detection. The method’s key addition was the usage of CNN which was improved by SBO, which improved the network’s accuracy in comparison to the traditional conjugate gradient approach.

To classify skin lesions, many researchers chose to conduct extensive experiments on the widely available skin cancer dataset named HAM10000 (Human Against Machine). In [10] the researchers initiated a simple transformer framework to classify skin cancer on this dataset incorporating a clinical dataset. For the HAM10000 dataset, in contrast with the state-of-the-art methods, the proposed model showed 94.3% accuracy. The multi-scaled network integrating contrastive learning showed the most efficient result having 94.1% accuracy which was much higher in comparison to the traditional CNN such as MobileNetV2, ResNet50, and InceptionV2. Aladhadh et al. (2022) developed a two-tier framework to overcome the challenges involved with accurately classifying skin diseases in another research [11] and presented a comprehensive evaluation of the MVT model on the HAM10000 dataset to classify SC images effectively. The accuracy of the training set was 98% and the validation set was 90% without any preprocessing of the data. The authors trained the same dataset using the GrabCut-stacked CNN (GC-SCNN) model with a fuzzy base [12]. Support vector machines (SVM) and fuzzy GC-SCNN in this case achieved almost 99.75% classification accuracy. Again, Wu et al. claimed that skin biopsy specimens are diagnosed by pathologists based on their visual assessments which often can be inaccurate [13]. To solve this problem, they initiated a Scale-Aware Transformer Network (ScAtNet) to categorize melanocytic skin lesions in digital WSI. To do this, CNN was utilized to separately develop patch-wise mappings to every data scale. Next, from the combined multi-scale patch embeddings, the model learned inner patch and inter-scale representation using a transformer. The network was compared with patch-based classification, ChikonMIL, weighted feature aggregation, streaming CNN, and MS-DA-MIL the test set achieved 64% accuracy for multiple input scales.

It is stated that in Medical Data Analysis, a crucial component for the efficacy of deep learning is a large-scale and thoroughly annotated dataset mentioned in [14]. However creating such massive annotations is quite difficult, especially for

histopathology pictures with distinctive features. Therefore, the researchers used the widest publicly accessible datasets of histopathology imaging for developing a transformer-based unsupervised feature extractor. The system's primary component was a combination of CNN and a multi-scale Swin Transformer architecture. The dataset in [7] was again trained using a CNN model. A trained VGG-16 model was used which showed better accuracy for sensitivity, NPV, and PPV. Furthermore, in paper [15] the authors assessed the potential of GANs (Generative Adversarial Networks) to address a variety of major issues related to cancer imaging, including data imbalance, domain, and dataset changes, access to data and confidentiality, data analysis and quantification, cancer determination, tumor profiling, and surgical planning. The literature on GANs used for cancer images was reviewed critically, and recommendations were offered for future research areas to solve these issues. Additionally, a feature-based responsive transformer network that utilizes the traditional decoder encoder architectural style, known as FAT-Net, was presented in [16] to completely regulate lengthy interdependence and broad relevant data. This network integrated an additional transformer subsidiary because standard CNN-based methods frequently struggle to achieve a satisfactory segmentation performance. This transformer encoder performed image segmentation utilizing a unique sequence-to-sequence predictive model, in contrast to typical CNN-based encoders.

The study [17] revealed that the processing of digital images could be used to classify thermal imaging pictures of skin cancer lesions. The researchers demonstrated that skin lesions that are cancerous possessed red component values exceeding 100 on a scale from 0-255 in the RGB color space. When segmenting based on the k-means algorithm thermography, the affected area's average value in the red component was greater than that in other areas including melanoma. Using these findings as a guide, a non-invasive tool for diagnosing skin cancer was created, cutting down on pointless diagnostic procedures and streamlining the diagnosis. The article [18] briefly discussed the utilization of computational Intelligence in Image Processing to detect skin cancer. It revealed the effectiveness of dermoscopy, the inspection of the pattern of the algorithms- the ABCD rule, the Menzies grading technique, and the seven-point checklist for assessing skin lesions. The primary aims of the study [19] were to assess several ViT architectures based on the recommended models and training techniques for breast cancer classification. The implemented Vision Transformer models' usefulness in the medical sector was proved as it outperformed ResNet-50 while using fewer model parameters. It is worth mentioning that in recent years, many academics have been working on creating computer-aided diagnostic (CAD) methods for classifying skin cancer [13]. Before the advent of deep learning, machine learning (ML) techniques were primarily employed in CAD systems. However, ML-based approaches can only identify a portion of skin illnesses because of the difficulty of feature building and the constraints of handmade features. Again, Deep Learning Algorithms are more accurate and effective at automatically learning semantic characteristics from large-scale datasets. Hence, the authors [20] proposed to improve computer-aided diagnostics by including non-standard picture decomposition techniques and using classification systems that utilize ensemble models with statistical learning in addition to conventional methods that focus on geometrical characteristics and color or pattern analysis. Their proposed methodology combined medical

expertise with multiple cutting-edge technologies including image processing, classification of patterns, statistical learning, and ensembling methods to classify skin lesions with the assistance of technology for diagnostic aid.

Skin lesion classification is a challenging aspect of dermatological image analysis and has seen notable progress with the integration of deep learning. The paper explores [21] the use of pre-trained Convolutional Neural Networks (CNNs) as feature extractors combined with other machine learning classifiers to improve skin lesion recognition accuracy. The study highlights the value of transfer learning by utilizing pre-trained models such as DenseNet201 and applying them to smaller datasets to increase efficiency. To ascertain their influence on categorization results, three data scenarios original, pre-processed, and augmented are investigated. Using datasets such as PH2 and ISIC 2019, the study examines 17 pre-trained CNN architectures and 24 classifiers to find viable combinations. The accuracy of DenseNet201 combined with Weighted KNN is noteworthy, achieving 62.43%, demonstrating the promise of this technique. The results additionally demonstrate the effectiveness of alternative combinations: ShuffleNet with Linear SVM (71.42%), DarkNet53 with LDA (67.29%), and ResNet18 with LDA (60.34%). In addition to highlighting deep learning’s versatility in dermatological image analysis, these findings provide insightful information for the development of skin lesion categorization systems in the future.

Using the same dataset, Nunnari et al. [22] investigate how to improve the classification of skin lesions in images by merging pixel data with patient metadata, such as age, gender, and body location. The study explores metadata fusion using shallow neural networks and non-neural machine learning techniques, VGG16 and RESNET50 as baseline Convolutional Neural Networks (CNNs). Per-class sensitivity declines for three of the four CNN cases, despite an overall accuracy improvement, hinting at possible difficulties for underrepresented classes. Surprisingly, seven out of 16 teams experienced decreased accuracy when integrating metadata, highlighting the practice’s limited acceptance in CNN-based architectures. The study emphasizes how crucial it is to assess the value of metadata and take dataset design biases into account. Model accuracy with VGG16 is 66.76%, and improvements are seen with VGG16+SVM (75.32%), VGG16+RF (76.79%), and VGG16+XGBoost (73.38%) which shows that in terms of improving accuracy, shallow neural networks do better than other methods. Moreover, EfficientNets, SENet, and ResNeXt WSL are the deep learning models combined in the paper[23] that were chosen using a search approach. Data-driven methods such as loss balancing, multi-crop evaluation, and an unknown class in the test set are used to overcome challenges such as severe class imbalance and moderate class imbalance. Making use of a variety of EfficientNets, the study emphasizes the importance of multi-resolution input and effective data augmentation. The addition of metadata enhances the method’s effectiveness, which helps smaller models in particular. The absence of metadata for the unknown class during training causes problems even with cross-validation improvements with more metadata. The research highlights the usefulness of different input resolutions and highlights how well EfficientNet performs in skin lesion classification.

Chapter 3

Research methodology

Figure 3.1, illustrating a general outline of the steps we intend to follow to conduct our research efficiently is given below:

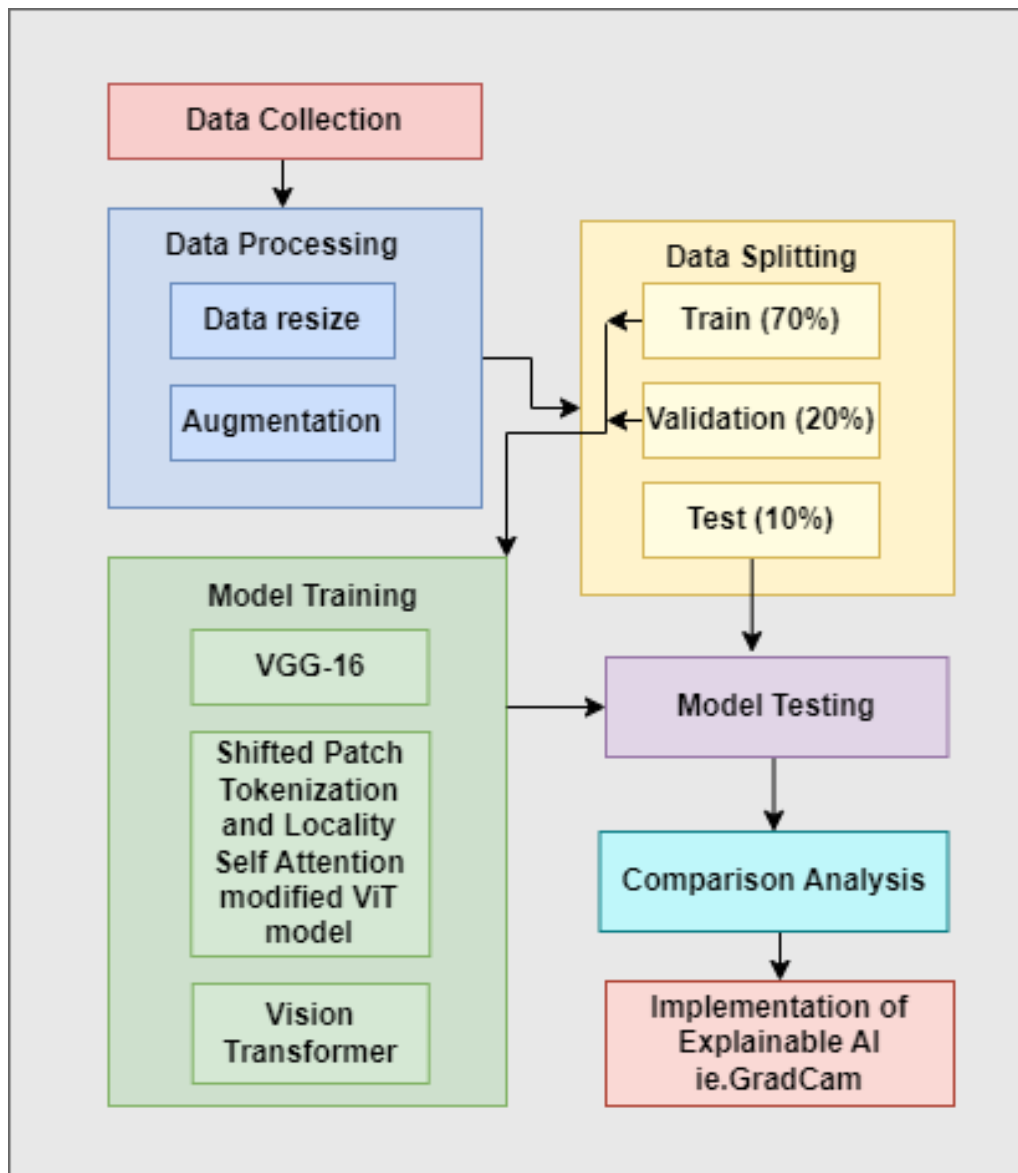


Figure 3.1: Our proposed methodology

For this research, we will use some skin cancer dataset. Following the collection of data, we will move to pre-processing. The techniques we intend to use to pre-process the data are image resizing and augmentation. Through image resizing technique, a uniform size will be chosen for all the images of the dataset whereas image augmentation involves altering the existing data to generate new training data for the model. After pre-processing is done, the data will be split into a 7:2:1 ratio which means training, validation, and test sets will include 70%, 20%, and 10% of data respectively. Following that, we will use the dataset to train our models. To observe the accuracy of the constructed model, we will utilize the test set from our dataset once it is trained. Next, we will analyze the performance of the models and then we will use the XAI technique on the best-performing model to find out the focal regions of the output images and analyze the results. Hence, our model will be able to detect cancerous skin lesions in the early stage of cancer.

3.1 Dataset

3.1.1 Dataset Source

The dataset we used for our work is ISIC2019 which is available on the ISIC archive [24]. The dataset contains a total of 25,331 images that are distributed among 8 distinct classes.

3.1.2 Description of the Dataset

The class-wise distribution of the images is shown below:

Class wise distribution			
Name of classes	Number of classes		
	Train	Test	Val
Actinic keratosis	716	75	76
Basal cell carcinoma	2820	250	253
Benign keratosis	2215	203	206
Dermatofibroma	206	11	22
Melanocytic nevus	10979	965	931
Melanoma	3812	360	350
Squamous cell carcinoma	541	42	45
Vascular lesion	202	24	27

Table 3.1: Class-wise data distribution

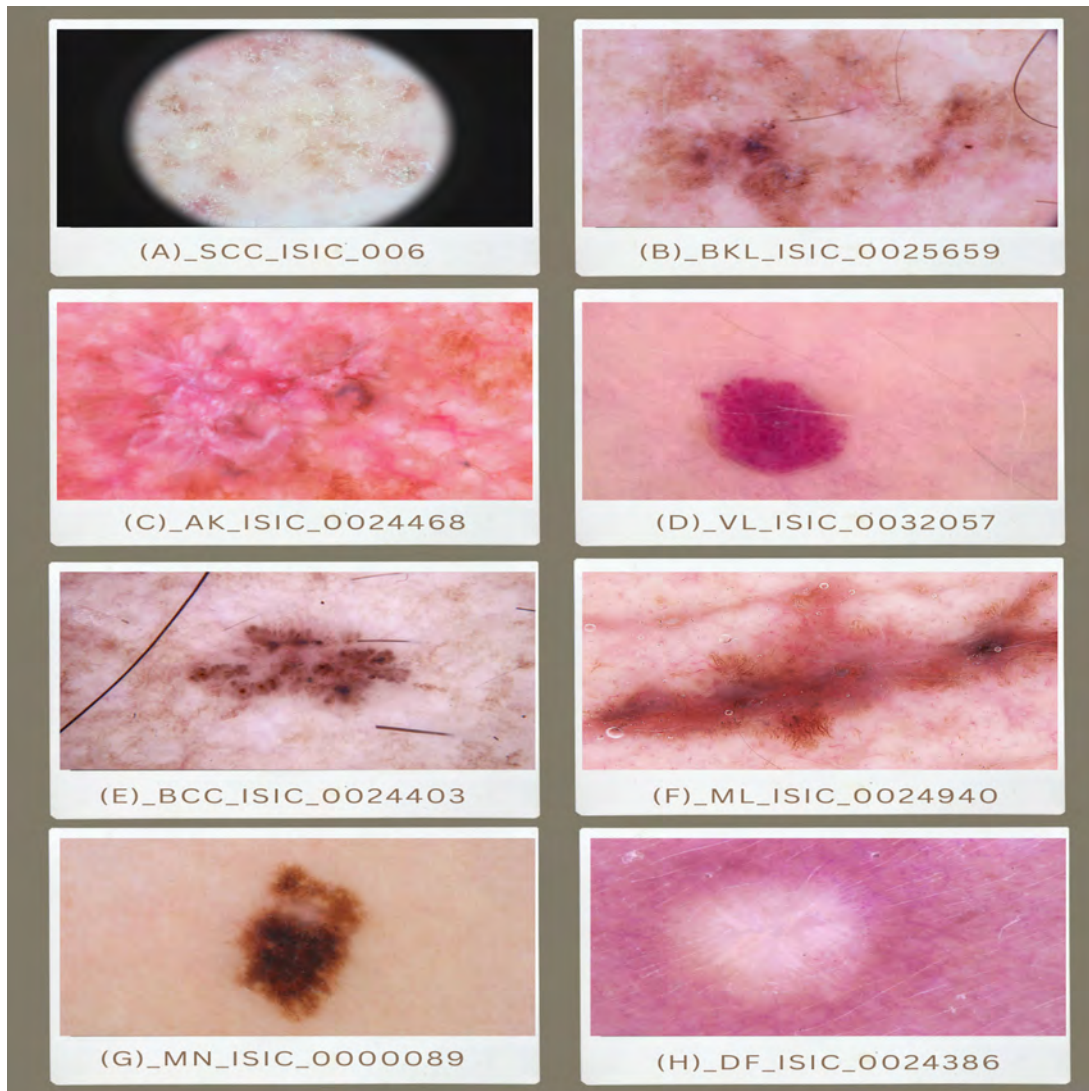


Figure 3.2: Sample of ISIC 2019 dataset: A) scc, B) bkl, C) akiec, D) vasc, E) bcc, F) mel, G) nv, and H) df

Melanoma is the deadliest type of skin cancer that occurs when melanocytes begin to proliferate uncontrollably. Its rapid growth and capability to spread to other organs make it more serious than other kinds of skin cancer. It can occur in any part of the body but is more likely to appear on the legs in the case of women and on the upper back of men. Melanomas are usually brown or black although they can also be red, pink, or flesh-toned.

Basal cell carcinoma (BCC) commonly develops when basal cells in the outermost layer of skin (epidermis) undergo alterations and grow uncontrollably due to the effects of ultraviolet (UV) radiation from the sun or indoor tanning. BCCs can have different appearances, including open sores, red spots, pink growths, shiny bumps, scars, and growths with slightly raised, rolling edges with a central indentation.

Small dry, scaly, or rough patches of skin are typical Actinic Keratosis appearances. They can be of any color—sometimes red, pink, light or dark tan, or flesh-toned and usually develop on the sun-exposed areas of the human body.

Squamous cell carcinoma (SCC) can be defined as the abnormal and rapid development of squamous cells. It typically occurs when ultraviolet radiation and other harmful factors cause aberrant alterations in the squamous cells. SCCs can take the form of scaly red patches, open sores, rough, thickened, or raised growths resembling warts on the scalp, ears, lips, or back of the hands.

Melanocytic nevus or mole is a prevalent and benign skin lesion caused by the localized enlargement of pigment-producing cells called melanocytes. It can be congenital or develop later in life. Melanocytic nevus can appear on any part of the body in mostly round or oval shapes and various colors from flesh tone to brown to black.

The fibrous tissue that makes up the dermis, the deeper of the two primary layers of skin, frequently overgrows, resulting in dermatofibroma. It is usually not cancerous and is benign. It can appear in pink to light brown color in light skin and dark brown to black color in dark skin.

Vascular lesions are clusters of blood vessel enlargements that are typically known as birthmarks. They can be categorized into three major types: Hemangiomas, the most common type which are benign tumors of blood vessel cells. The other two types are vascular malformations which refer to congenital abnormalities in vascular development and capillary malformations which are caused by the uneven capillary and tiny veins of the skin's deeper layer.

Benign keratosis is a general term to define similar scaly skin lesions like Solar lentigo, Seborrheic keratosis, and Lichen planus-like keratosis. Solar lentigo refers to macular hyperpigmented lesions caused by prolonged exposure to sunlight. They can be oval, round, or uneven in shape and mostly appear in light brown. Seborrheic keratosis is a noncancerous growth of keratinocytes in the epidermis, generally prevalent among aged people. Except for the mucous membranes, soles, and palms, they can appear on any area of the body. Another frequent benign skin growth is Lichen planus-like keratosis which is either a regressing seborrheic keratosis or solar lentigo. The features of such lesions are pink, brown, gray, and black colored papules or macules.

While working on the dataset, we observed that the dataset has a very imbalanced distribution. The highest number of instances(10979) exist in the 'Melanocytic Nevus' class. Compared to that, classes like 'Actinic keratosis', 'Dermatofibroma', 'Squamous cell carcinoma', and 'Vascular Lesions' have a very negligible amount of instances which are 716, 206, 541, and 202 images respectively. Therefore, we disregarded those four classes and utilized the images of the remaining four classes to train and evaluate the models.

3.2 Data Preprocessing

3.2.1 Data Resize

The ISIC2019 dataset is a combination of HAM10000, BCN20000, and MSK datasets. The HAM10000 dataset consists of images with a dimension of $600 * 450$ whereas the BCN20000 has images of $1024 * 1024$ dimensions. On the other hand, the MSK dataset contains images of various pixel sizes. To address the issue of various image sizes, the data has been processed through resizing. So, all our input images in the training dataset were resized into $128 * 128$ pixels.

3.2.2 Data Augmentation

A further step of data augmentation has been implemented to address the unequal distribution of the images across the classes. By applying techniques like random rotation, zooming, horizontal/vertical flipping, and normalization on the existing training samples, new data were generated for the classes for further use.

3.3 Data Splitting

As the dataset was highly imbalanced, we decided to work with the four classes that had the highest number of instances among all the classes. We took 1800 instances from each of these classes and split them into training, validation, and testing sets.

3.3.1 Training Set

To make the model learn potential patterns and hidden features of the data, the training set is fed into the model. For training, we have used 70% of the data which means 1260 images from each class have been used to train the model.

3.3.2 Validation Set

The validation dataset is used to adjust the hyperparameters of a classifier. In our case, the validation set consists of 20% data which means 360 images from each class are used to validate the performance of our model at the time of training.

3.3.3 Testing Set

A test dataset is a sample that assesses the model's fit without any bias, which can be used to estimate the model's accuracy. Among the 1800 images, 180 images, which is 10% of the distribution, are allotted for our model's testing.

3.4 Used Architectures

Below is a brief explanation of the CNN and transformer models that we will be using in this research:

1. VGG-16

The VGG-16 model is a deep convolutional neural network with 16 layers which is the initial for the Oxford Visual Geometry Group [18]. The VGG model that we used follows a transfer learning approach by initializing the weights of the original VGG-16 of a pre-trained model from imageNet. It is a 16-layer model with 13 convolutional layers. The convolutional layers have 3 * 3 filters with a stride length of 1 which are followed by 2 * 2 max pooling layers each with a stride length of 2. The model has 3 fully connected layers and 2 dropout layers. The convolutional layers are connected with the GeLU activation function and the output layer is connected with the softmax activation function.

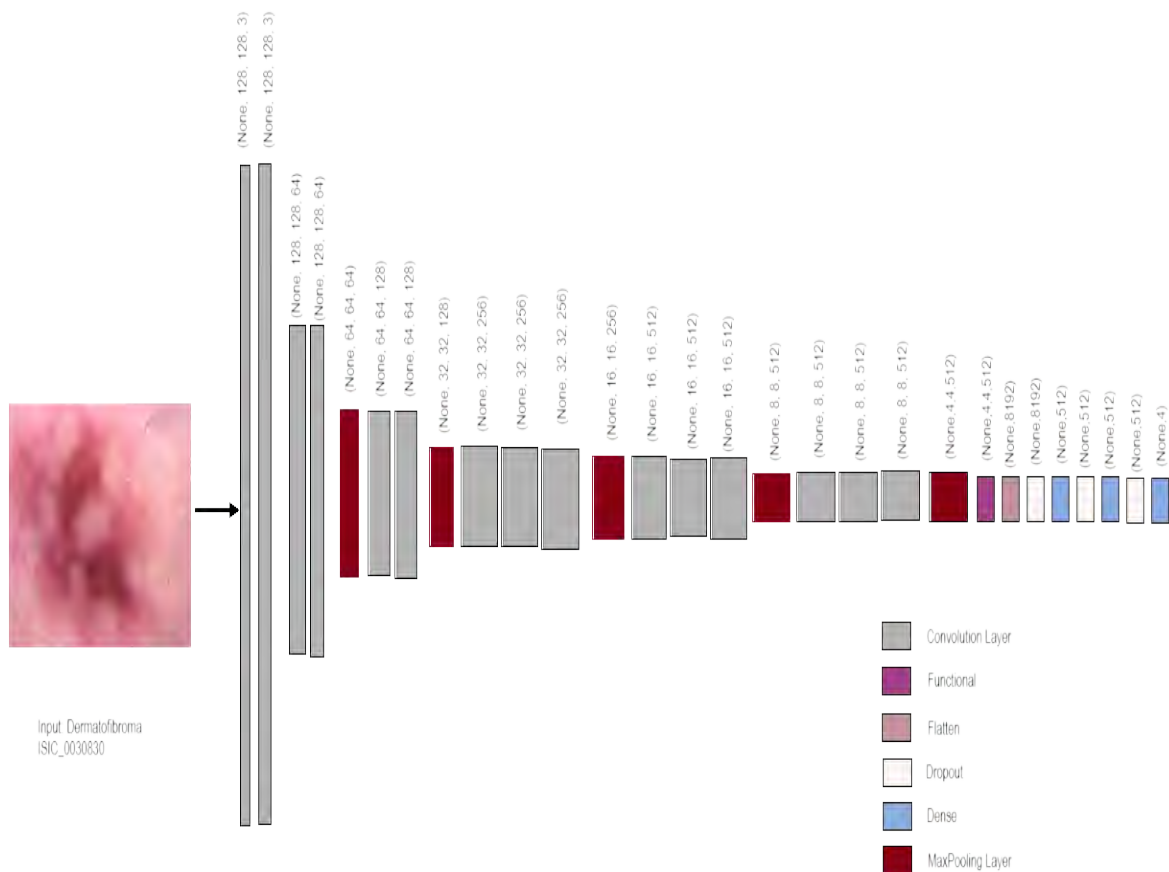


Figure 3.3: Internal Architecture of VGG-16

2. Vision Transformer (ViT)

A Vision Transformer is a transformer designed specifically for vision processing applications such as image identification. The vision transformer concept in computer vision uses multiple units of layers without the application of bias tailored to particular images. The performance of a vision transformer model is determined by decisions such as the optimizer, network length, and specific dataset hyperparameters [25]. The ViT models work by dividing input images into $16 * 16$ patches that are flattened, linearly projected in a dimension of 32, and converted into lower-dimensional embeddings. These embeddings then go through a transformer encoder of size $64 * 32$, followed by 4 Multilayer Perceptron (MLP) heads of size $1024 * 512$ with 8 fully connected layers to classify images.

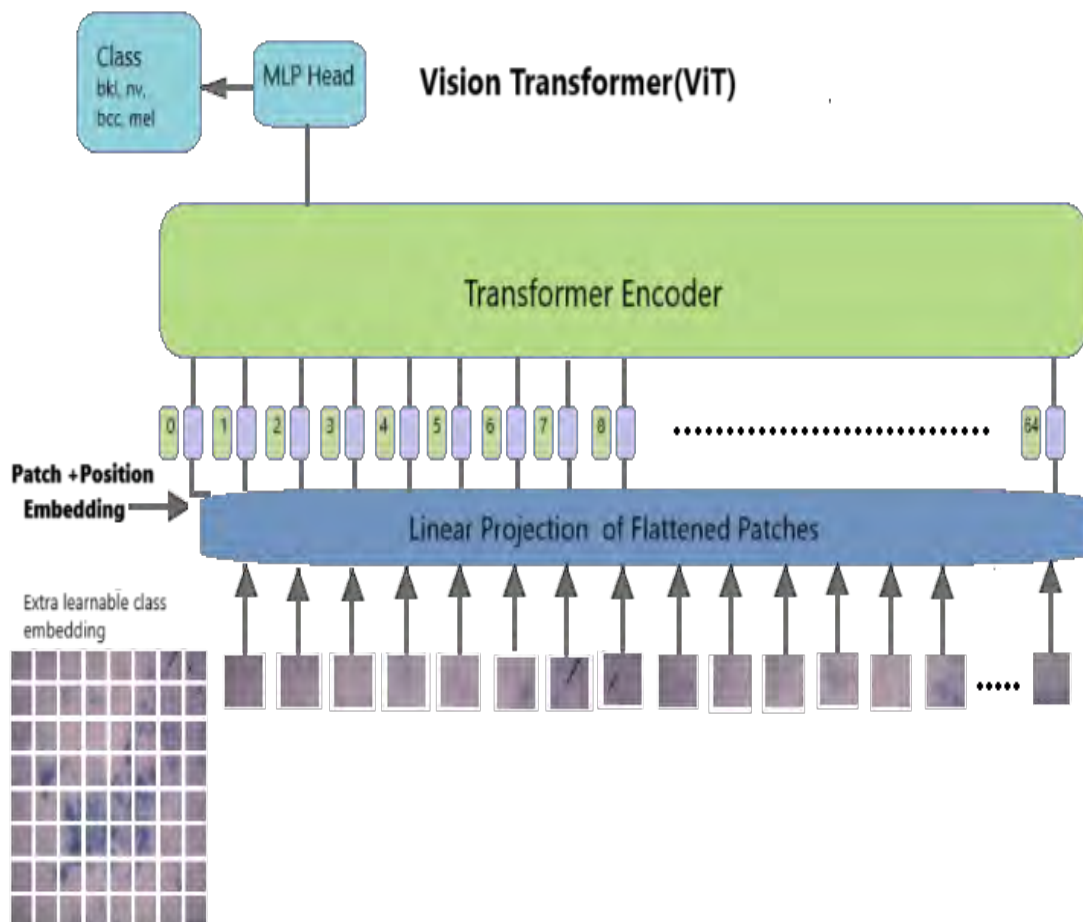


Figure 3.4: Internal Architecture of Vision Transformer

3. Shifted Patch Tokenization and Locality Self Attention modified ViT model (ViT-STP)

Since its inception in 2020, Vision Transformers (ViTs) have undergone various enhancements. One notable modification involves integrating Shifted Patch Tokenization (SPT) and Locality Self Attention (LSA). This alteration allows ViT to focus on local correlations between $128 * 128$ image pixels, reducing the reliance on larger datasets. Unlike the normal ViT model, SPT introduces a spatial shift to images in four diagonal directions up-left, up-right, down-left, and down-right. This shift improves the way how input images are represented spatially which merges shifted features into a single image [26]. After images undergo spatial shifting, they are split into non-overlapping patches and concatenated with the model input which improves the model's performance [27]. These overlapping patches are linearly embedded into vectors of dimension 32, creating the initial tokenized input for the transformer, where positional embeddings are added. In Shifted Patch Tokenization, a crucial modification occurs in the tokenization process. Rather than treating patches individually, adjacent patches are combined to form tokens representing shifted pairs. This shift aids in capturing relationships between neighboring patches that were flattened and transformed into lower-dimensional embeddings. These embeddings then undergo a transformer encoder of size $128 * 64$, followed by 4 MLP heads of size $2048 * 1024$ with 8 fully connected layers.

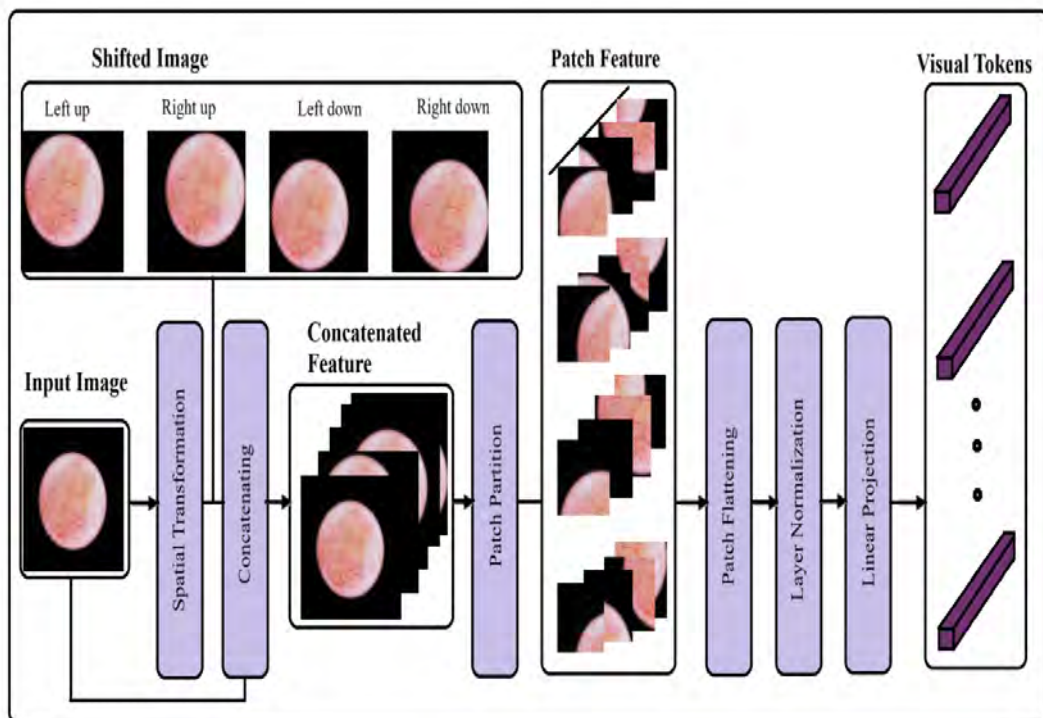


Figure 3.5: Internal Architecture of Shifted Patch Tokenization and Locality Self Attention modified ViT model (ViT-STP)

Chapter 4

Implementation And Result Analysis

4.1 Experimental Setup

All the experiments for this research have been performed on AMD Ryzen 5 5600X 6 Core processor, and a single GPU (Zotac GeForce GTX 1660 AMP Edition). Tensorflow with Keras, the deep learning framework is used to build the models. The dimensions of the training images are 128 * 128 and the batch size is 32. AdamW has been used as an optimizer. The learning rate is 0.00001 and the training epoch is set to 100. To ensure a fair comparison, we maintained identical experimental setups and overall architectures for the models. Several evaluation matrices are implemented to assess the performance such as accuracy, precision, recall, F1 score, and so on.

4.1.1 Confusion Matrix

Confusion matrices are $N \times N$ matrices employed to estimate the performance of classification models, where N determines the total target classes. These matrices help us to see how effectively our classification model performs and what sorts of errors it makes by comparing the actual values with the model's predicted values.

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	TN
	Negative	FN	FP

Table 4.1: Confusion Matrix

The above is shown as a representation of a 2×2 confusion matrix where the columns stand for the actual values and the rows stand for the predicted values of the target variable. Both true positive (TP) as well as true negative (TN) denotes that the predicted value is the same as the actual value. When the actual and the predicted values are positive, it falls under the true positive (TP). On the other hand, when both the values are negative, it falls under true negative (TN). False

negative (FN) and false positive (FP) imply false predictions of the model. When the model predicts a negative value for an actual positive value, is referred to false negative (FN) whereas the model predicting a positive value for an actual negative value is called a false positive (FP).

4.1.2 Accuracy

A model's accuracy is generally defined as the overall performance of the model across all classes of data. This metric is particularly efficient for balanced classification tasks. It is directly connected with the values of the confusion matrix. Accuracy can be calculated in the following way:

$$\text{Accuracy} = \text{Number of correct predictions} / \text{Number of total predictions}$$

4.1.3 Precision

Precision measures the ability of the model to classify positive samples correctly while considering the negative samples. Precision is determined by dividing the number of accurately segregated positive samples by the sum of all the Positive samples (both correct and incorrect).

$$\text{Precision} = \text{True Positive} / (\text{True Positive} + \text{False Positive})$$

4.1.4 Recall

Recall also determines the model's potential to classify positive samples but does not take negative samples into account while classifying. A higher value of recall indicates that more positive cases have been detected by the model. The calculation of sensitivity goes as follows:

$$\text{Recall} = \text{True Positive} / (\text{True Positive} + \text{False Negative})$$

4.1.5 F1 score

F1 score is a better-suited measure for solving imbalanced classification problems. It represents the harmonic mean of precision and recall. The value typically ranges from 0 to 1. An F1 score closer to 1 means the model has a lower false positive and lower false negative. When the value is 1, means the model is perfect.

$$\text{F1 Score} = 2 * ((\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}))$$

4.2 Performance Measure

The performance measures of all the used models are as follows:

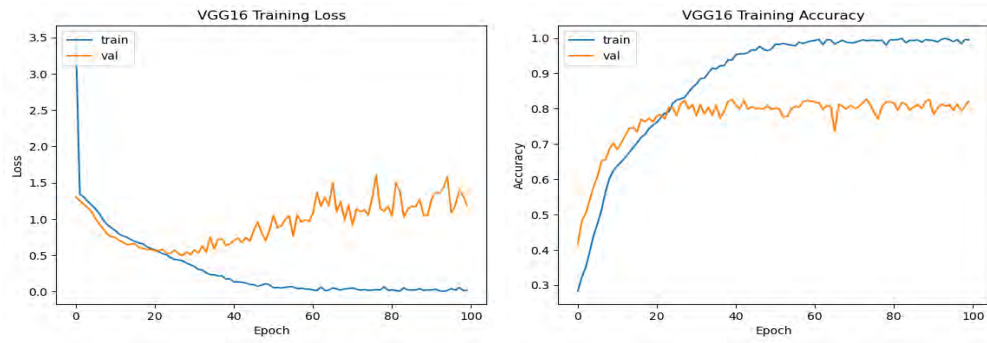


Figure 4.1: Loss and accuracy of VGG-16 model

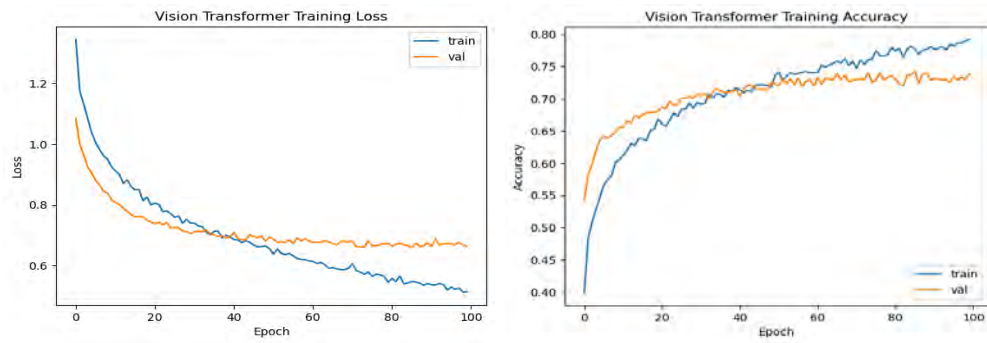


Figure 4.2: Loss and accuracy of ViT model

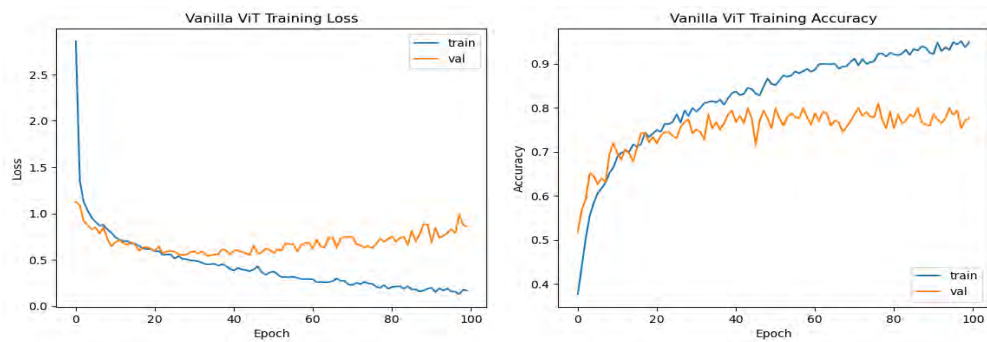


Figure 4.3: Loss and accuracy of vanilla ViT model



Figure 4.4: Loss and accuracy of Shifted Patch Tokenization and Locality Self Attention modified ViT model

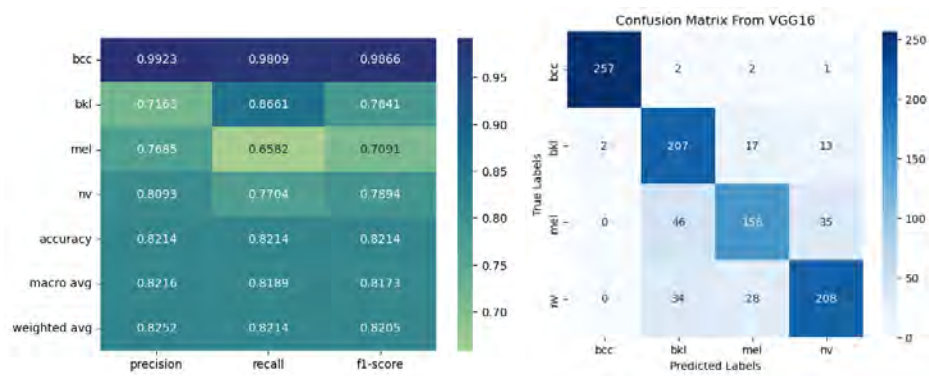


Figure 4.5: Classification report and confusion matrix of VGG-16 model

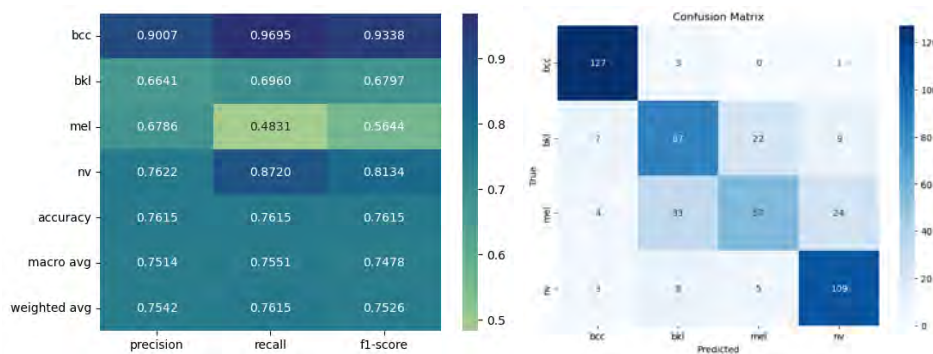


Figure 4.6: Classification report and confusion matrix of ViT model

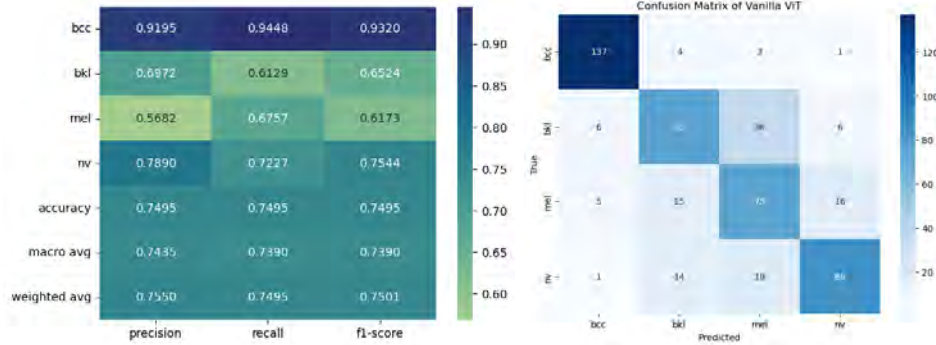


Figure 4.7: Classification report and confusion matrix of vanilla ViT model

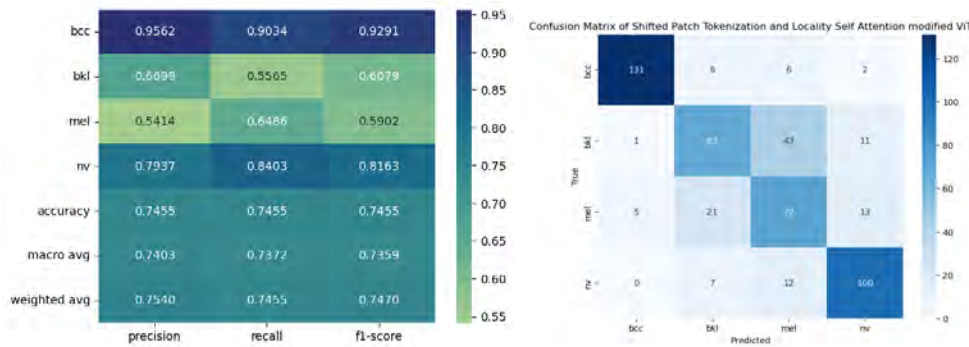


Figure 4.8: Classification report and confusion matrix of Shifted Patch Tokenization and Locality Self Attention modified ViT model

4.3 Result Analysis

In Table 4.2, there is a detailed analysis of our results with other research papers' results using the same dataset, ISIC 2019. In the paper of Nunnari et al. [22], using the VGG-16 model on the dataset, they achieved an accuracy of 66.76% whereas our accuracy reached 82.14%. The reason for this difference in accuracy mainly lies in the data augmentation and configuration settings chosen during the model training. Their paper indicates that the model was trained only on a subset of images from 8 imbalanced classes of the dataset for 10 epochs. To avoid training bias and achieve a better performance score, we balanced the data of the 4 largest classes and trained the model for 100 epochs. They used a batch size of 8 with an SGD optimizer. On the other hand, our model was trained on a batch size of 32. Due to the AdamW optimizer's competency in deeper CNN models [28], we used it as an optimizer over SGD. Hence, in our case, the VGG16 model showed better performance on the dataset. While comparing the performance of VGG16 with the vision transformer, we found that ViT showed slightly reduced performance scores than VGG16. The possible reason behind this performance drop could be that ViTs need a very large amount of data [29] to be trained properly. We further compared the original ViT with the shifted patch tokenization and locality self-attention modified ViT which generally tends to improve the performance of ViT [30] for smaller-size datasets. However, it was observed that the difference between the performance scores of the original and the modified ViT is very small and the inclusion of the modified ViT has not conferred significant benefits in our scenario. Therefore, we can conclude that the VGG16 model has outperformed the other transformer models.

Study	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Gessert 2020 [23]	Ensembles of multi resolution EfficientNet	63	-	73	-
Nunnari et al. 2020 [22]	VGG16	66.76	-	65.97	55.97
	VGG16+ GradBoost	59.20	-	52.36	42.93
	VGG16+SVM	75.32	-	65.43	64.63
	VGG16+RF	76.79	-	67.28	67.54
	VGG16+ XGBoost	73.38	-	67.30	64.20
Beniyahia et al. 2021 [21]	DenseNet201+ Weighted KNN	62.43	74.47	-	58.11
	ResNet18+LDA	60.34	-	-	-
	ShuffleNet+ Linear SVM	71.42	-	-	-
	EfficientNet B0+ GNB	40	-	-	-
	DarkNet53+LDA	67.29	-	-	-
Our used models	VGG16	82.14	82.52	82.14	82.05
	VGG16+GradCam	-	-	-	-
	ViT	76.15	75.42	76.15	75.26
	ViT+ Shifted Patch	74.95	75.50	74.95	75.01
		74.55	75.40	74.55	74.70

Table 4.2: Result Analysis

Chapter 5

GradCam

It is seen that the model's architecture and parameters can be easily visualized at a detailed level, but this does not provide us with a complete view of what the model perceives when it solves a specific issue. But it can be viewed using Explainable or Interpretable Artificial Intelligence (XAI) which is the idea of providing a human-readable explanation for a model's decision-making process [31].

Within the realm of XAI, GradCam is a popular method that helps to bridge the gap between the high-dimensional representation of the models and the human-understandable representation of the models. The full form of GradCam is Gradient-weighted Class Activating Mapping. Usually, in deep learning architecture, a class-specific technique is used to understand the focal regions of an input image which are important for the prediction of the network of a particular class while making a decision. GradCam produces an output that is essentially a heatmap visualization for a particular class label which can be used to visually confirm the pixels in the image which is been focused by the model [32]. There are several benefits of GradCam. They are:

1. Comprehending models' performance,
2. Assisting the cognitive comprehension of the model's functionality, and
3. Providing visual information regarding the model's emphasis.

Moreover, the output of GradCam shows which particular pixels of the images are required by the model for the performance of the classification tasks. There are two steps for the implementation of GradCam:

1. Identifying the last convolutional layer in the network, and
2. Observing the gradient information flowing into the last layer.

For our case, we found that our VGG-16 model was performing better so we decided to use GradCam on the VGG-16 model. In the initial implementation of GradCam, we had to identify the last convolutional layer which is "block5-conv3", also load the checkpoint which is the h5 file of the model, and observe the gradient information flow in the last layer. After implementing GradCam on the VGG-16 model some output images were generated which are as follows:

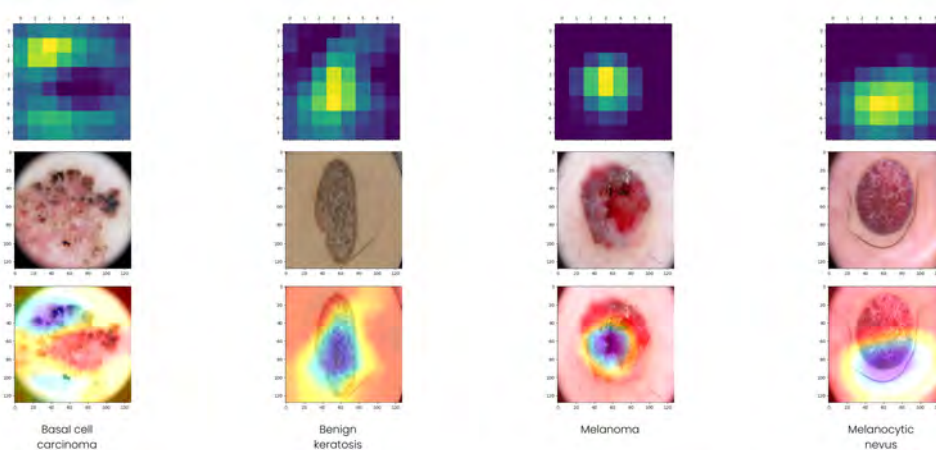


Figure 5.1: GradCam outputs for 4 different classes (BCC, BKL, MEL, NV) on VGG-16 model

In GradCam when the input image is passed through the network layer it produces a final classification score. After that, it calculates the gradient of the prediction with the feature maps of the last convolutional layer which is done using backpropagation. Then Global Average Pooling (GAP) is done to obtain the corresponding weight to each important feature map. Subsequently, weighted sum, ReLU, and Upsampling are done on the predicted class and finally, a heatmap is generated which is our final result of the GradCam. Here, the heatmap highlights the region that plays a role in the contribution of the model's prediction. The heatmap has several classes of colormaps where we used the "miscellaneous" class [33].

In Figure 4.1, for all the classes it is observed that in the region where the cancerous cell is located, the heatmap generates a darker gradient meaning that the model focuses on that region while making a decision and our model performed well in the detection of skin cancer cells and it gave an accurate decision. To verify the detection and classification of skin lesions by our model, it can be compared with the traditional method which is Dermatoscopy. In the clinical classification of BCC, it is detected by seeing the arborizing telangiectasias, large blue/gray oval nest-like nodules, ulceration, leaf-like regions, and structureless reddish white background [34]. In above Figure 5.1, we can see that the visualization of the activation mapping is densely shown in the regions where there is swelling, large blue/gray oval nest-like nodules and in the rest of the region, the gradient is sparse meaning that the model did not focus on those regions while making its decision.

Chapter 6

Conclusion

It is seen that early detection of skin cancer can help to cure it. So for this reason other than Dermatoscopy deep learning models can be implemented to detect it. CNN models like VGG-16 have long dominated the area of medical image analysis, but sometimes they are confined to focusing on local alterations in visual patterns. Due to the amazing performance and expanding potential of employing transformer processes, Vision Transformers are gradually becoming the next emerging trend in computer vision. In the world of health, even an erroneous classification result might endanger lives. Therefore, we can not fully rely on the model's prediction and for this reason, we incorporated explainable AI (XAI) to interpret the outcomes of the model and help in the accurate decision-making process. It is seen that all the machine learning models are like black-boxes meaning we can not see the intricate internal architecture of the models. All the deep learning models have several hidden layers that make the internal logic of the network quite difficult to understand for human beings. While the parameters and architecture of the network can easily be viewed at a technical level it does not fully provide the complete visualization of what the model does while doing certain tasks. To make it easier to understand our model and come to a conclusion, we used Gradcam on the model to observe the focal region of our model in detecting skin lesions and making a decision. We discovered that in our case, the VGG-16 model outperformed the other models with an accuracy of 82.14% which is why we used GradCam on it to understand the internal function of whether the model can perform well and make an accurate decision by detecting the correct region that is responsible for classifying the cancerous cells.

Bibliography

- [1] C. Xin, Z. Liu, K. Zhao, *et al.*, “An improved transformer network for skin cancer classification,” *Computers in Biology and Medicine*, vol. 149, p. 105 939, 2022.
- [2] Y. Gulzar and S. A. Khan, “Skin lesion segmentation based on vision transformers and convolutional neural networks—a comparative study,” *Applied Sciences*, vol. 12, no. 12, p. 5990, 2022.
- [3] H. M. Balaha and A. E.-S. Hassan, “Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm,” *Neural Computing and Applications*, pp. 1–39, 2022.
- [4] Y. Gulzar and S. A. Khan, “Skin lesion segmentation based on vision transformers and convolutional neural networks—a comparative study,” *Applied Sciences*, vol. 12, no. 12, p. 5990, 2022.
- [5] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, “Skin cancer classification with deep learning: A systematic review,” *Frontiers in Oncology*, vol. 12, 2022.
- [6] Z. Xu, F. R. Sheykhahmad, N. Ghadimi, and N. Razmjoooy, “Computer-aided diagnosis of skin cancer based on soft computing techniques,” *Open Medicine*, vol. 15, no. 1, pp. 860–871, 2020.
- [7] S. Aneja, N. Aneja, P. E. Abas, and A. G. Naim, “Transfer learning for cancer diagnosis in histopathological images,” *arXiv preprint arXiv:2112.15523*, 2021.
- [8] R. Mohakud and R. Dash, “Designing a grey wolf optimization based hyperparameter optimized convolutional neural network classifier for skin cancer detection,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 6280–6291, 2022.
- [9] M. Ogorzałek, G. Surówka, L. Nowak, and C. Merkwirth, “New approaches for computer-assisted skin cancer diagnosis,” in *The Third International Symposium on Optimization and Systems Biology, Zhangjiajie, China, 2009*.
- [10] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, and M. Islam, “An effective skin cancer classification mechanism via medical vision transformer,” *Sensors*, vol. 22, no. 11, p. 4008, 2022.
- [11] U. Bhimavarapu and G. Battineni, “Skin lesion analysis for melanoma detection using the novel deep learning model fuzzy gc-scnn,” in *Healthcare*, MDPI, vol. 10, 2022, p. 962.
- [12] R. Osuala, K. Kushibar, L. Garrucho, *et al.*, “Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging,” *Medical Image Analysis*, p. 102 704, 2022.

- [13] H. Dong and F. Farid, “A deep learning based patient care application for skin cancer detection,” 2022.
- [14] K. Thomsen, A. L. Christensen, L. Iversen, H. B. Lomholt, and O. Winther, “Deep learning for diagnostic binary classification of multiple-lesion skin diseases,” *Frontiers in medicine*, vol. 7, p. 574 329, 2020.
- [15] N. Paliwal, “Skin cancer segmentation, detection and classification using hybrid image processing technique,” *International Journal of Engineering and Applied Sciences*, vol. 3, no. 4, p. 257 678, 2016.
- [16] M. Ogorzałek, G. Surówka, L. Nowak, and C. Merkwirth, “Computational intelligence and image processing methods for applications in skin cancer diagnosis,” in *International Joint Conference on Biomedical Engineering Systems and Technologies*, Springer, 2010, pp. 3–20.
- [17] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, “Fat-net: Feature adaptive transformers for automated skin lesion segmentation,” *Medical Image Analysis*, vol. 76, p. 102 327, 2022.
- [18] J. O. Emuoyibofarhe, D. Ajisafe, R. S. Babatunde, and M. Christoph, “Early skin cancer detection using deep convolutional neural networks on mobile smartphone,” *International Journal of Information Engineering & Electronic Business*, vol. 12, no. 2, 2020.
- [19] J. Daghrir, L. Tlig, M. Bouchouicha, and M. Sayadi, “Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach,” in *2020 5th international conference on advanced technologies for signal and image processing (ATSIP)*, IEEE, 2020, pp. 1–5.
- [20] J. Lindroos, “Transformers for breast cancer classification,” 2022.
- [21] S. Benyahia, B. Meftah, and O. Lézoray, “Multi-features extraction based on deep learning for skin lesion classification,” *Tissue and Cell*, vol. 74, p. 101 701, 2022.
- [22] F. Nunnari, C. Bhuvaneshwara, A. O. Ezema, and D. Sonntag, “A study on the fusion of pixels and patient metadata in cnn-based classification of skin lesion images,” in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, 2020, pp. 191–208.
- [23] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin lesion classification using ensembles of multi-resolution efficientnets with meta data,” *MethodsX*, vol. 7, p. 100 864, 2020.
- [24] [Online]. Available: <https://www.isic-archive.com/>.
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [26] S. H. Lee, S. Lee, and B. C. Song, “Vision transformer for small-size datasets,” *arXiv preprint arXiv:2112.13492*, 2021.
- [27] A. Emmamuel, U. Asim, H. Yu, S. Kim, *et al.*, “3d-cnn method over shifted patch tokenization for mri-based diagnosis of alzheimer’s disease using segmented hippocampus,” *Journal of Multimedia Information System*, vol. 9, no. 4, pp. 245–252, 2022.

- [28] E. M. Dogo, O. J. Afolabi, and B. Twala, “On the relative impact of optimizers on convolutional neural networks with varying depth and width for image classification,” *Applied Sciences*, vol. 12, no. 23, p. 11 976, 2022.
- [29] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, “How to train your vit? data, augmentation, and regularization in vision transformers,” *arXiv preprint arXiv:2106.10270*, 2021.
- [30] S. Lee, S. Lee, and B. C. Song, “Improving vision transformers to learn small-size dataset from scratch,” *IEEE Access*, vol. 10, pp. 123 212–123 224, 2022.
- [31] H. Zhang and K. Ogasawara, “Grad-cam-based explainable artificial intelligence related to medical text processing,” *Bioengineering*, vol. 10, no. 9, p. 1070, 2023.
- [32] V. Jahmunah, E. Y. K. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, “Explainable detection of myocardial infarction using deep learning models with grad-cam technique on ecg signals,” *Computers in Biology and Medicine*, vol. 146, p. 105 550, 2022.
- [33] [Online]. Available: <https://matplotlib.org/stable/users/explain/colors/colormaps.html>.
- [34] M. C. Fargnoli, D. Kostaki, A. Piccioni, T. Micantonio, and K. Peris, “Dermoscopy in the diagnosis and management of non-melanoma skin cancers,” *European Journal of Dermatology*, vol. 22, no. 4, pp. 456–463, 2012.