Comparative Analysis of Convolutional Neural Network and vision Transformer based algorithms for Diagnosing pigmented skin lesions

Aswathi P

Research Proposal

November 2023

# Abstract

The forthcoming study aims to delve into the evolving role of Transformers within the computer vision domain, specifically exploring their potential to potentially supplant or challenge Convolutional Neural Networks (CNNs) in identifying pigmented skin lesions. Through a comparative study, the study will seek to elucidate the viability and efficacy of Vision Transformers (ViTs) by pitting three state-of-the-art ViT models against a widely recognized CNN architecture. The research aims to reveal the surprising revelation that the ViT is expected to outperform the CNN significantly. It is anticipated that the ViT will showcase markedly higher accuracy levels and F1-scores across its class classifier, firmly establishing its prowess in this specialized domain.

This study embarks on a comprehensive exploration into the evolving landscape of computer vision, particularly honing in on the realm of identifying pigmented skin lesions. Its primary objective revolves around a comparative analysis between two formidable architectures: Convolutional Neural Networks (CNNs) and the emerging Vision Transformers (ViTs). This comparative study aims to ascertain whether ViTs possess the potential to supplant or challenge the longstanding dominance of CNNs in the specialized domain of pigmented skin lesion identification. By leveraging three state-of-the-art ViT models and juxtaposing their performance against a widely recognized CNN architecture, this research endeavors to reveal potential paradigm shifts in the field of computer vision. The anticipated findings of this study are poised to illuminate the anticipated superiority of Vision Transformers over Convolutional Neural Networks in the realm of recognizing pigmented skin lesions. It is expected that ViTs will showcase a notable leap in performance metrics, displaying significantly higher accuracy levels and F1-scores across class-specific classification tasks. Notably, these findings suggest that ViTs exhibit a propensity for enhanced performance with augmented training data, a notable characteristic that sets them apart from the more traditional CNN model. The emergence of this observation alludes to the tantalizing prospect of Transformers potentially replacing CNNs in this particular task, underscoring their transformative potential and marking a potential shift in the landscape of computer vision methodologies applied to medical diagnostics and skin cancer identification.

**Table of Contents**

[Abstract 1](#_Toc540073940)

[1. Background 3](#_Toc1875144151)

[2. Problem Statement 3](#_Toc1419144339)

[2.1. Related Research 4](#_Toc1326904296)

[3. Research Questions (If any) 9](#_Toc120059588)

[4. Aim and Objectives 9](#_Toc2118924152)

[5. Significance of the Study 9](#_Toc1391581492)

[6. Scope of the Study 10](#_Toc403345701)

[7. Research Methodology 11](#_Toc632748113)

[8. Requirements Resources 17](#_Toc1916374999)

[9. Research Plan 18](#_Toc1075603204)

[References 18](#_Toc1209101706)

# Background

Convolutional Neural Networks (CNNs) originated in the 1960s and 1970s, evolving over time through multiple researchers' contributions to become the cornerstone of modern computer vision. These networks have revolutionized image classification, object detection, segmentation, and medical imaging, with ongoing research focused on enhancing their interpretability, resilience, and versatility across various domains. In healthcare, CNNs significantly impact disease diagnosis by analyzing complex medical images like X-rays, MRIs, and CT scans, enabling early detection and contributing to precision medicine through tailored treatments and surgical guidance. Their specialized algorithms excel in identifying intricate patterns within images, aiding in swift and accurate anomaly detection, while their ability to segment specific regions proves crucial for tasks like radiation therapy planning, showcasing their pivotal role in advancing medical diagnostics and treatments.

The Vision Transformer (ViT) burst onto the computer vision scene as a recent breakthrough, unveiled in Dosovitskiy et al.'s pivotal paper "An Image Is Worth 16x16 Words: Transformers for Image Recognition" in 2020. In a departure from traditional Convolutional Neural Networks (CNNs), ViT embraced the Transformer architecture, initially renowned in natural language processing (NLP). Instead of relying on CNNs' methodology, ViT employs self-attention mechanisms, processing images as sequences of tokens, akin to how transformers handle textual sequences. Its outstanding performance in image classification challenged CNNs' dominance, proving its mettle across diverse datasets and showing competitiveness on benchmarks like ImageNet. ViT's advent spurred interest in harnessing transformer-based models for medical imaging tasks, paving the way for ongoing research and advancements. These strides in ViT and similar transformer-based architectures hold the promise of significantly elevating the accuracy, efficiency, and precision of medical image analysis, promising improved patient care and outcomes within healthcare settings.

# 2. Problem Statement

Since 2020, there has been an emerging recognition of transformers, particularly Vision Transformers (ViTs), showcasing their potential to rival or even surpass CNNs in image classification tasks. The successful integration of the attention mechanism—initially prominent in natural language processing—into computer vision has notably expanded transformers' applicability to image-related functions. However, in the realm of diagnostic applications for pigmented skin cancer, predominant research in the past three years has concentrated heavily on exploring CNNs' capabilities in predicting various skin lesion types. Researchers have experimented extensively with diverse pre-trained deep models such as VGG19, Darknet-53, NasNet-mobile, and InceptionResNetV2, while some studies have fine-tuned deep CNNs through specialized preprocessing to enhance skin lesion prediction accuracy. Despite the limited exploration of vision transformers in this domain, relying solely on accuracy as the primary evaluation metric might inadequately represent the transformers' effectiveness in diagnostic tasks. Consequently, this research seeks to introduce the F1-score as an additional metric to comprehensively compare the performance of two crucial neural network architectures—CNNs and ViTs—while harnessing transfer learning techniques.

The investigation acknowledges transformers' ability to achieve comparable or superior outcomes to CNNs in exploring the viability and effectiveness of cutting-edge ViT models for dermoscopy image classification tasks. Although the chosen CNN model for comparison lacks complexity or diversity, the evident impact achieved by ViT remains undeniable. Looking ahead, to further enhance the transformer's capacity in recognizing pigmented skin lesions, augmenting the training data appears pivotal. This could be achieved by introducing more manually labelled images or exploring the use of generative adversarial neural networks to generate additional data, thus aiming to achieve a more balanced dataset.

## **2.1. Related Research**

Here are the synopses and findings from the related papers.

Convolution Neural Networks (CNNs):

“Very Deep Convolutional Networks for Large-Scale Image Recognition” (Simonyan & Zisserman, 2014) research examines how convolutional network depth influences accuracy in large-scale image recognition. It focuses on assessing networks with increasing depth, using a unique architecture with small (3x3) convolution filters. By expanding the network to 16–19 layers, we observed a significant performance improvement, crucial to our success in the 2014 ImageNet Challenge where we achieved top rankings in both localization and classification. Our study highlights the versatility of our representations across different datasets, consistently delivering cutting-edge results. To encourage further exploration in utilizing deep visual representations in computer vision, we've made our powerful ConvNet models available to the public

The paper “Going deeper with convolutions “(Szegedy et al., 2014) a revolutionary convolutional neural network architecture celebrated for its strides in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14) in classification and detection tasks. This architecture stands out for its ability to maximize computational resources by skillfully balancing depth, width, and consistent computational costs. Emphasizing quality optimization, the design draws from the Hebbian principle and multi-scale processing intuition. The paper specifically spotlights GoogLeNet, a 22-layer deep network, highlighting its exceptional quality and performance in both classification and detection during ILSVRC14.

Convolutional networks are pivotal in modern computer vision, advancing benchmarks significantly since 2014 with remarkably deep models. While larger models boost quality with ample labeled data, efficient computation and fewer parameters are crucial for applications like mobile vision and big-data scenarios. In “Rethinking the Inception Architecture for Computer Vision “(Szegedy et al., 2015) study delves into efficient network scaling, emphasizing factorized convolutions and strict regularization. Evaluation on the ILSVRC 2012 classification challenge's validation set reveals substantial progress, achieving 21.2% top-1 and 5.6% top-5 error rates with a network totaling 5 billion multiply-adds per inference and fewer than 25 million parameters. Employing a 4-model ensemble and multi-crop evaluation further showcases remarkable results, reporting 3.5% top-5 error and 17.3% top-1 error.

Training deeper neural networks poses significant challenges. In “Deep Residual Learning for Image Recognition” (He et al., 2015) a novel residual learning framework to overcome challenges in training deeper networks than previous models. Their method involves redefining layers to learn residual functions regarding the layer inputs, leading to optimized training and improved accuracy in deeper residual networks. Evaluation on ImageNet included depths up to 152 layers, surpassing VGG nets while maintaining lower complexity. The ensemble achieved a remarkable 3.57% error rate on the ImageNet test set, securing 1st place in ILSVRC 2015. Additionally, assessments with 100 and 1000 layers on CIFAR-10 highlighted the importance of representation depth in visual recognition, showcasing a 28% improvement on the COCO object detection dataset. These deep residual nets significantly contributed to 1st place rankings in ImageNet detection, localization, COCO detection, and segmentation in ILSVRC & COCO 2015 competitions.

A powerful convolutional neural network successfully classified 1.2 million high-resolution images into 1000 distinct classes during the ImageNet LSVRC-2010 contest in “ImageNet Classification with Deep Convolutional Neural Networks” (Krizhevsky et al., n.d.) study. The test data revealed impressive performance with top-1 and top-5 error rates of 37.5% and 17.0%, respectively, surpassing prior benchmarks. This neural network comprises five convolutional layers, some followed by max-pooling layers, and three fully-connected layers culminating in a 1000-way softmax output, boasting 60 million parameters and 650,000 neurons. To expedite training, non-saturating neurons and an efficient GPU implementation for convolutions were utilized. Employing "dropout," a potent regularization method, countered overfitting in the fully connected layers. In the ILSVRC-2012 competition, this model excelled with a top-5 test error rate of 15.3%, notably surpassing the second-best entry at 26.2%.

Vision Transformer (ViT):

The Transformer architecture, popular in natural language tasks, sees limited use in computer vision, often integrated with convolutional networks or selectively employed to substitute certain components while retaining the network's structure. In “An image is worth 16x16 words: Transformers for image recognition at scale” (Dosovitskiy et al., 2020) research questions the exclusive reliance on CNNs and showcases a pure transformer's efficacy when directly applied to sequences of image patches for high-performance image classification tasks. ViT (Vision Transformer), pre-trained on extensive datasets and tested on mid-sized or small image recognition benchmarks like ImageNet, CIFAR-100, and VTAB, outperforms state-of-the-art convolutional networks while demanding notably reduced computational resources during training.

Newly, attention-based neural networks proved effective for image understanding, particularly in image classification, yet these high-performing vision transformers often demand pre-training with vast datasets comprising millions of images and significant computational resources, limiting their broader use. In “Training data-efficient image transformers & distillation through attention” (Touvron et al., 2020) study introduces competitive convolution-free transformers trained solely on Imagenet in less than 3 days on a single computer. The leading vision transformer (86M parameters) achieves an impressive 83.1% top-1 accuracy on ImageNet without external data. It introduces a teacher-student strategy tailored for transformers, leveraging a distillation token enabling the student to learn via attention mechanisms from the teacher. This token-based distillation, especially when using a convnet as a teacher, yields competitive outcomes akin to convnets for ImageNet (reaching up to 85.2% accuracy) and transfer learning scenarios.

The Vision Transformer (ViT), a recent addition to vision tasks, segments images into token sequences using multiple Transformer layers to grasp their global relationships. However, ViT falls short compared to CNNs on midsize datasets like ImageNet due to its tokenization method's simplicity, hindering the capture of vital local structures and inefficient training sample utilization, along with its attention-heavy architecture constraining feature richness within fixed computation budgets and limited training samples. To address these challenges, a novel Tokens-To-Token Vision Transformer (T2T-ViT) is proposed, in “Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet” (Yuan et al., 2021) study integrates a layer-wise Tokens-to-Token (T2T) transformation to structure images progressively into tokens, improving the modeling of local structures and reducing token length. This optimized T2T-ViT backbone, inspired by CNN design principles, impressively reduces parameters and operations by 50% compared to ViT, resulting in over a 3.0% performance boost on ImageNet from scratch. Surpassing ResNets and matching MobileNets' performance when directly trained on ImageNet, T2T-ViT, similar in size to ResNet50 (21.5M parameters), achieves outstanding 83.3% top-1 accuracy at 384×384 image resolution on ImageNet.

This paper “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows” (Liu et al., 2021) adapting the Transformer architecture for various computer vision tasks. Overcoming challenges of scale differences among visual elements and higher image pixel resolutions than text words, it employs a hierarchical structure and Shifted windows for efficient representation computation. This innovation optimizes computational efficiency by confining self-attention within local non-overlapping windows while enabling connections across them. The Swin Transformer excels in image classification, object detection, and semantic segmentation, outperforming prior benchmarks and showcasing the potential of Transformer-based models in vision tasks.

The recent advancements of Transformer models in computer vision have motivated researchers to enhance the original Pyramid Vision Transformer (PVT v1). In this study “PVT v2: Improved Baselines with Pyramid Vision Transformer” (Wang et al., 2021), three key modifications were introduced to PVT v1, namely (1) a linear complexity attention layer, (2) overlapping patch embedding, and (3) a convolutional feed-forward network. Through these augmentations, PVT v2 successfully decreases the computational complexity of its predecessor to a linear scale while demonstrating significant performance enhancements across core vision tasks like classification, detection, and segmentation. Impressively, the proposed PVT v2 achieves performance levels comparable to or even surpassing recent state-of-the-art models like the Swin Transformer. This endeavour aims to foster further advancements in Transformer-based research within the realm of computer vision.

Skin lesions data related papers:

The report of “Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions” (Díaz, 2017) they participated in the ISIC 2017 Challenge on Skin Lesion Analysis for Melanoma Detection, particularly in Lesion Classification. Their submission aimed to automate diagnosing nevus, melanoma, and seborrheic keratosis by merging dermatological expertise with Convolutional Neural Networks (CNNs), renowned for visual recognition. Their strategy focused on identifying lesion areas, segmenting structural patterns, and providing clinical diagnoses. Additionally, they introduced specialized CNN blocks to integrate dermatological insights into the diagnostic process.

This article “Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)” (Codella et al., 2017) the dermoscopic image analysis benchmark challenge aimed to improve melanoma automated diagnosis, featuring tasks like lesion segmentation, feature detection, and disease classification. The event drew 593 registrations, 46 finalized submissions, and around 50 attendees, making it a comprehensive and comparative study. Although rankings concluded, the challenge dataset remains open for ongoing research and development.

This study “Data Augmentation for Skin Lesion Analysis” (Perez et al., 2018) delves into the effectiveness of 13 different data augmentation techniques for melanoma classification using three CNNs (Inception-v4, ResNet, and DenseNet). The study explores diverse augmentation methods, including color shifts, geometric transformations, elastic transforms, random erasing, and lesion blending, demonstrating their significant impact on training and testing. The most effective scenario reaches an AUC of 0.882 for melanoma classification, outperforming the top submission (0.874) in the ISIC Challenge 2017 that utilized extra data for training.

Detecting skin cancer early significantly improves treatment outcomes, but annotated data for training AI models is often limited and expensive to acquire. To overcome this challenge, the study “Skin Lesion Synthesis with Generative Adversarial Networks” (Bissoto et al., 2019)   
Using Generative Adversarial Networks (GANs), we aim to generate convincing synthetic skin lesion images, addressing data scarcity by creating clinically relevant and visually authentic data, potentially enhancing automated skin cancer classification systems beyond current capabilities.

This paper “Skin Lesion Classification Using Ensembles of Multi-Resolution EfficientNets with Meta Data”(Gessert et al., 2019) details our approach for the ISIC 2019 Skin Lesion Classification Challenge involving dermoscopic images, we tackled two tasks: lesion classification and utilizing patient metadata. Our approach involved a data-driven method leveraging external data to cover missing skin lesion types, addressing class imbalance with loss balancing techniques, and handling image resolution variations. Incorporating metadata, we merged additional neural networks' features with CNN and optimized models through ensembling, achieving top rankings in both tasks with a balanced accuracy of 63.6% for task 1 and 63.4% for task 2 on the official test set.

# 3. Research Questions

Is it possible to attain equivalent accuracy between CNNs and ViTs with the same dataset, or does ViT demonstrate the potential to achieve higher accuracy compared to CNNs using identical data?

Is it feasible to achieve an identical inference speed for both CNNs and ViTs when processing the same input?

Can Vision Transformer models outperform CNNs in the task of diagnosing pigmented skin lesions, and how do both types of models respond to increases in training data?

# 4. Aim and Objectives

**Aim:** To evaluate the feasibility and effectiveness of Vision Transformers (ViTs) in identifying different types of pigmented skin lesions from dermoscopy images and compare their performance to Convolutional Neural Networks (CNNs) in the context of skin cancer diagnosis.

**Objectives:**

1. To compare the performance of ViTs and a customized CNN model in classifying pigmented skin lesions from dermoscopy images.
2. To assess the overall accuracy and F1-score per class of the classifier for both ViTs and CNN.
3. To determine if increasing the amount of training data for ViTs can further improve their performance.
4. To investigate whether ViTs can potentially replace CNNs for this specific task if a larger and balanced dataset of dermoscopy images becomes available.

These objectives align with the study's goal of assessing the potential of ViTs as an alternative to CNNs in improving the diagnosis of skin diseases from dermoscopy images.

# 5. Significance of the Study

The significance of this study is multi-faceted and impactful:Advancing Understanding of Vision Transformers (ViTs) in Dermatology: Performance Comparison: The study compares ViTs against a prominent Convolutional Neural Network (CNN) in identifying pigmented skin lesions. The ViT's surprising outperformance signifies a potential paradigm shift in leveraging transformer architectures for dermatological diagnoses.

Potential Replacement of CNNs: Transformers' Superiority: The ViT's notably higher accuracy and F1-scores per lesion class compared to the CNN hint at the possibility of ViTs surpassing and potentially replacing CNNs in this specific diagnostic task.

Implications for Data Volume and Model Performance: Data Impact: Observations regarding ViT's response to increased training data contrasted with the CNN's stability shed light on the importance of dataset size. The potential for ViT's performance enhancement with more data suggests scalability and adaptability.

Future Directions in Dermatological Diagnosis: Impact on Clinical Practices: If validated and implemented, the ViT's superiority could revolutionize dermatological diagnosis. Improved accuracy could enhance early detection and prognosis, significantly impacting patient outcomes.

Guiding Further Research and Development - Influence on AI Research: The study's findings may guide future research toward ViTs in medical imaging tasks. Insights gained could spur innovations in neural network architectures and training methodologies for dermatological diagnostics.

Paving the Way for Enhanced Diagnostic Tools - Potential for Advanced Diagnostic Systems: ViTs' demonstrated potential suggests the development of more accurate and efficient diagnostic systems. These systems could benefit healthcare by aiding dermatologists and potentially reducing the burden on healthcare resources.

# 6. Scope of the Study

The study's scope is to delve into Vision Transformers' effectiveness in pigmented skin lesion classification, specifically focusing on their performance with dermoscopy images. It aims to scrutinize varying training data sizes and diverse evaluation metrics to contribute towards enhanced diagnostic capabilities in dermatology.

The investigation primarily centres on classifying pigmented skin lesions sourced from dermoscopy images, pivotal for early detection and diagnosing various skin conditions, notably skin cancer. These images offer detailed skin analysis and serve as the foundation for automated classification, essential for dermatologists' practice.

The research involves exploring Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) models—DeiT, Swin Transformer, and CrossViT—to evaluate their efficacy in lesion classification. Leveraging transfer learning and fine-tuning techniques, the study adapts pre-trained models to augment performance for this specific task.

Performance evaluation encompasses metrics like overall accuracy and F1-score, especially critical due to dataset imbalances. Utilizing the extensive HAM10000 dataset labeled with various skin diseases diagnosed through medical methods, the study examines the impact of different training data sizes and divisions on model adaptability. Tailoring optimization strategies, learning rate schedules, and loss functions specific to each model category (CNN and ViT), the research strives for optimal outcomes. Ultimately, it aims to contribute to dermatology by enhancing automated pigmented skin lesion classification, potentially improving dermatologists' clinical practices.

In essence, this study's comprehensive scope is to thoroughly explore the ViT models' applicability in pigmented skin lesion classification using dermoscopy images. It seeks to gauge their performance in comparison to CNN models and offer insights into the potential advantages of integrating ViT models into dermatological practices.

# 7. Research Methodology

**Model Selection:** The study utilizes two categories of models: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The selected Vision Transformer (ViT) models to be explored in the study will include DeiT (Data-efficient image Transformer), Swin Transformer, and CrossViT. Each of these models will bring distinct innovations and modifications to the foundational ViT structure:

DeiT (Data-efficient image Transformer) (Touvron et al., 2020): Distillation Token: DeiT will include an additional token termed the distillation token. This will enhance the model's performance by allowing it to learn from a teacher model through distillation.

Swin Transformer (Liu et al., 2021): Hierarchical Feature Maps: Swin Transformer will introduce the concept of hierarchical feature maps, allowing the model to capture information at various scales.

Shifted Window Attention: Instead of the standard self-attention mechanism, Swin Transformer will employ shifted window attention. This approach will enhance efficiency by focusing attention computation on non-overlapping local windows.

CrossViT: Dual Branches: This will use separate dual branches to extract multi-scale features, enhancing its ability to capture diverse information. Cross-Attention Module: This module will be utilized to merge the features extracted from the dual branches, enabling CrossViT to effectively process and combine multi-scale information.

The selection process for the Convolutional Neural Network (CNN) models will involve picking architectures with parameters exceeding 85 million, intending to explore their capabilities for a specific task:

Parameter Threshold: Models with a parameter count surpassing 85 million will be considered.

* Architectural Complexity: Emphasis will be placed on architectures that demonstrate complexity through a deep stack of layers and wide network widths.
* Identification: Architectures such as DenseNet, EfficientNet, or ResNet variants with higher complexities and substantial parameter counts will be included.
* Layer Configurations: The selected CNN models will consist of a series of convolutional layers, residual connections, pooling operations, and potentially include additional components like bottleneck blocks or dense connections.

**Model Architecture:** ResNet9: The architecture of ResNet9, Figure 1 is designed in a hierarchical manner comprising three main stages, each contributing to the model's depth and complexity. The initial two stages follow a similar pattern, featuring two consecutive convolutional blocks. These blocks are pivotal for extracting intricate features from the input data. Subsequently, an identity block is introduced after each convolutional block, which employs two convolutional layers. These identity blocks play a role in refining the learned representations and enhancing the model's feature extraction capabilities.

Moving to the final stage, a series of operations are executed to consolidate the learned features. This stage incorporates essential processes such as max pooling, which condenses the learned information, followed by flattening to prepare the data for subsequent layers. Dropout layers are integrated into this stage to mitigate overfitting by randomly deactivating certain neurons during training. Ultimately, this stage concludes with a dense layer, which refines the accumulated information to align with the number of classes involved in the classification task.

The ResNet9 model, built upon this three-stage structure, introduces three different variant models. These variations are established upon the fundamental framework of the initial transformer-like architecture tailored explicitly for image classification. By leveraging the foundational design of this transformer-inspired architecture, ResNet9 implements modifications and adaptations to further optimize its performance and adaptability in classifying images across various datasets and tasks.

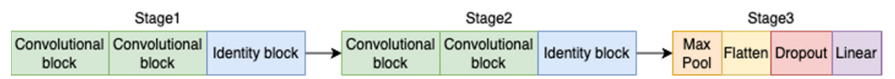


Figure 1: ResNet9 (Yang, 2023)

If we consider ResNet-9 Figure 2 as the architectural backbone, it will be chosen for its simplicity and strong performance in image classification tasks, composed of 9 layers with skip connections mitigating gradient vanishing. This approach will involve the training of two hemispheres: specializing the left on specific classes and the right on general classes, with the objective of extracting distinct yet overlapping features. These hemispheres will then merge into a bilateral architecture, combining their penultimate layer outputs and incorporating two heads dedicated to specific and general classes. After the initial training phase, the weights of individual hemispheres will be frozen to preserve learned features, while focus will shift to training the newly integrated heads.

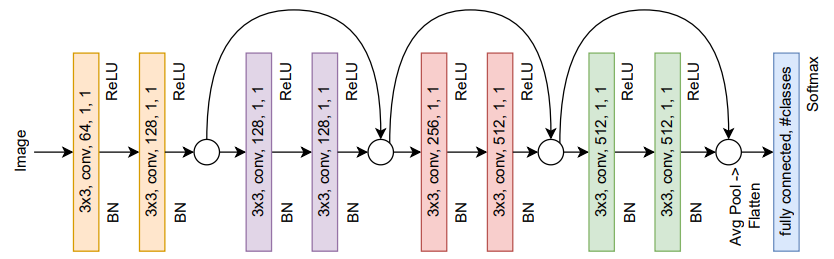


Figure 2: ResNet-9 architecture: A convolutional neural net with 9 layers and skip connections (Rawlinson et al., n.d.)

ViT: The fundamental architecture of the Vision Transformer (ViT), Figure 3 typically encompasses three core stages, each contributing to the overall process of transforming images into meaningful classifications.

Stage 1 initiates the image processing pipeline by breaking down the input images into smaller patches. This patch-wise decomposition of the image allows for efficient handling of visual information. These patches, once extracted, are then flattened to prepare them for further processing. By flattening the patches, the spatial information contained within these smaller segments is organized into a format suitable for subsequent computational steps.

Stage 2 involves the transformation of these patches into embedded representations. This transformation occurs through a combination of linear embeddings and positional embeddings. Linear embeddings serve to convert the flattened patches into feature vectors with semantic representations, effectively capturing the essential characteristics of the patches. Simultaneously, positional embeddings encode spatial information, ensuring that the model understands the relative positioning of these patches within the image. These embedded patches form the basis for the subsequent stages of processing.

Stage 3 focuses on leveraging the embedded patch representations to predict the classes or categories associated with the input images. The embedded patches are fed into a transformer encoder, a crucial component that allows ViT to process and analyze the visual information hierarchically and capture complex relationships between patches. Additionally, an extra multi-layer perceptron (MLP) linear layer is employed, serving as a final processing step that refines the extracted features before the classification decision is made. This MLP layer aids in enhancing the model's ability to discriminate between different classes based on the learned representations from the transformer encoder.

Ultimately, by progressing through these three stages, the ViT framework systematically transforms image data into a structured format that can be effectively processed and classified by utilizing a combination of patch-based representations, positional information, and transformer-based encoding mechanisms.

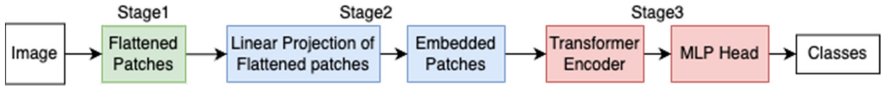


Figure 3: ViT (Yang, 2023)

The approach Figure 4 involves breaking down an image into consistent patches, applying linear embedding to each, integrating position embeddings, and then passing this sequence through a standard Transformer encoder. For classification purposes, a learnable "classification token" is included within the sequence. The design of the Transformer encoder draws from Vaswani et al.'s work in 2017.

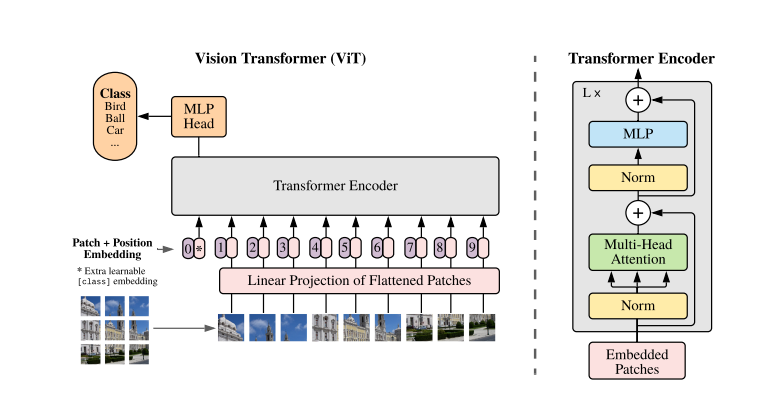


Figure 4: Model overview (Dosovitskiy et al., 2020)

**Data Preprocessing:** Data augmentation techniques play a pivotal role in enhancing the quality and diversity of training datasets, especially in image-related tasks such as image classification. These techniques encompass a variety of transformations applied to the training data to augment its variability and aid in the robustness and generalization of machine learning models. Among these techniques, random flips, which horizontally or vertically mirror images, contribute to introducing different orientations, thereby enriching the dataset with varied perspectives. Additionally, adjustments to brightness, contrast, saturation, and hue within images emulate diverse lighting conditions and colour distributions, enabling the model to adapt to varying visual environments. Random erasure of small image patches strategically obscures parts of images, promoting the learning of robust features by preventing the model from overfitting to specific details. Simultaneously, data normalization, standardizing the data to a uniform scale or distribution, ensures consistency and aids in model convergence by offering a standardized representation across the dataset's features. Together, these augmentation methods collectively bolster the training data, equipping machine learning models to better handle diverse scenarios and perform adeptly when presented with unseen data during inference or testing phases.

**Model Training**: The training process for Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) models diverges in their optimization strategies, schedulers, and loss functions.

ResNet9, a CNN architecture, opts for an Adam optimizer during gradient descent, which is a popular optimization algorithm used in deep learning. Alongside this optimizer, ResNet9 employs a one-cycle learning rate (LR) scheduler, which dynamically adjusts the learning rate during training. Additionally, it utilizes a cross-entropy loss function, a standard choice for classification tasks, effectively measuring the difference between predicted and actual class probabilities.

In contrast, ViT models utilize different optimization strategies. They employ an AdamW optimizer, a variant of Adam that introduces weight decay, enhancing its stability and performance. Coupled with this optimizer, ViT models employ a cosine LR scheduler, which smoothly decreases the learning rate as training progresses, aiding in fine-tuning and convergence. Moreover, ViT models utilize a label smoothing cross-entropy loss function, which incorporates a smoothing term into the standard cross-entropy loss. Label smoothing is a regularization technique commonly used in sophisticated image classification models. It prevents the model from becoming overconfident by slightly adjusting the ground truth labels, thereby enhancing generalization and improving model calibration.

This disparity in optimization techniques between CNNs and ViT models underscores their distinct approaches to addressing image classification tasks. While both aim for accurate predictions, their divergent strategies reflect their unique design principles and considerations in handling and processing visual data.

**Evaluation Metrics:** The evaluation matrix in this research extends beyond overall accuracy by incorporating the F1-score measurement to assess the quality of each classifier. This choice stems from the inherent imbalance within the HAM10000 dataset. The F1-score, known for its effectiveness in evaluating classifiers within unbalanced datasets or when prediction errors hold significant consequences, becomes particularly relevant in this context. Moreover, considering the practical implications of deploying skin cancer classifiers, inference speed emerges as an additional evaluation metric. The ability to process predictions swiftly is crucial in real-time clinical settings, enhancing the classifiers' usability and practicality. Hence, accuracy, F1-score, and inference speed collectively serve as essential metrics for comprehensively evaluating the performance and applicability of the classifiers.

**Data Set:** The HAM10000 dataset is a widely used collection of dermatoscopic images for research in skin lesion classification. It comprises various attributes and metadata associated with each image. The data dictionary for HAM10000 typically includes:

1. Image ID: Unique identifier for each image in the dataset.
2. Lesion ID: Identifier for the specific lesion or skin condition represented in the image.
3. Image: The actual dermoscopic image of the skin lesion.
4. Diagnosis: The clinical diagnosis or label assigned to the skin lesion (e.g., melanoma, nevus, basal cell carcinoma, etc.).
5. Age: Age of the patient when the image was taken.
6. Sex: Gender of the patient.
7. Anatomical Site: The location on the body where the lesion is present (e.g., back, abdomen, arm, etc.).
8. Localization: Specific region on the anatomical site (e.g., upper, lower, left, right, etc.).
9. Dataset Split: Some datasets include a split identifier for training, validation, or testing purposes.
10. Other Metadata: Additional information such as the date the image was taken, patient details, or further clinical annotations.

This data dictionary provides a comprehensive overview of the various attributes associated with each image in the HAM10000 dataset, aiding researchers in understanding and analysing the characteristics of skin lesions and facilitating the development of classification models and algorithms.

**Research Focus:** The primary focus of this study is to compare the efficacy of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in accurately classifying pigmented skin lesions within dermoscopy images. Additionally, the research aims to investigate how increased training data impacts the performance of Vision Transformers. This investigation centers on evaluating the distinct capabilities of CNNs and ViTs in handling the classification task while exploring how ViTs respond to larger training datasets. The study aims to highlight the strengths and limitations of both architectures in dermatological image classification, shedding light on ViTs' adaptability and performance under varying data volumes.

# 8. Requirements Resources

The study on dermoscopy image classification using Vision Transformers (ViT) and Convolutional Neural Networks (CNN) requires various resources and data. Here are some of the key requirements and resources:

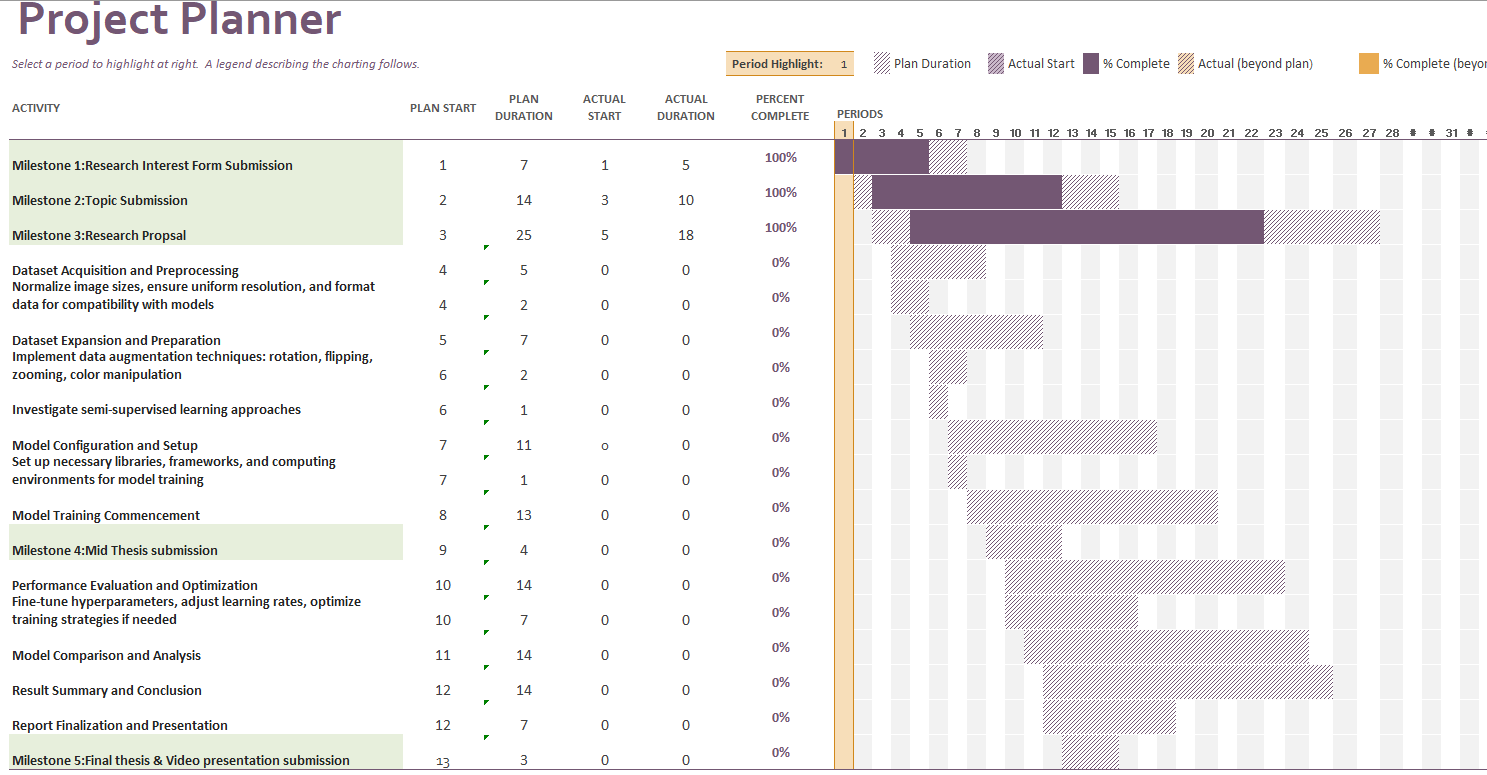
Hardware Resources:

* CPU: A multi-core processor (quad-core or higher) for data processing and model training.
* GPU (Recommended, but not mandatory): For faster training, especially with larger models or datasets, a dedicated GPU is beneficial, At least an NVIDIA GeForce GTX 1050 Ti or higher for moderate tasks, For more computationally intensive tasks, consider higher-end GPUs like NVIDIA RTX series or Quadro GPUs.
* RAM: A minimum of 8GB RAM is advisable for basic experimentation, for larger datasets or complex models, 16GB or more RAM is recommended.
* Storage: Adequate disk space to store the dataset and model checkpoints. SSDs are preferred over HDDs for faster data access.

Software Resources:

* Pandas (1.3+): For handling structured data efficiently.
* NumPy (1.20+): Numerical computations and scientific computing.
* scikit-learn (0.24+): Tool for data preprocessing, model evaluation
* OpenCV (4.5+): Tools for image and video analysis
* TensorFlow (2.3+): Widely used for building neural networks and machine learning models.
* PyTorch (1.5+): ResNet implementations in the torchvision.models module.
* Keras (standalone or integrated with TensorFlow 2.3+): for ResNet model access.
* torchvision (for PyTorch 1.9+): specifically tailored for computer vision tasks.
* TensorFlow Datasets (TFDS 4.4+): Facilitating dataset management within TensorFlow.
* Matplotlib (3.4+): A comprehensive plotting library
* Seaborn (0.11+): for high-level interfaces drawing attractive & statistical graphics.
* TensorBoard (for TensorFlow 2.6+): tracking and visualizing model training TensorFlow.
* tqdm (4.62+): adds progress bars to loops and tasks-visual representation of progress
* h5py (3.3+): Allows interaction with datasets stored in the HDF5 file format
* Pickle: serialization module used for object serialization and deserialization

# 9. Research Plan

Figure 5: Gantt chart planner

# References

1. Bissoto, A., Perez, F., Valle, E., & Avila, S. (2019). *Skin Lesion Synthesis with Generative Adversarial Networks*. <https://doi.org/10.1007/978-3-030-01201-4_32>
2. Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2017). *Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)*. <http://arxiv.org/abs/1710.05006>
3. Díaz, I. G. (2017). *Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions*. <http://arxiv.org/abs/1703.01976>
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. <http://arxiv.org/abs/2010.11929>
5. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2019). *Skin Lesion Classification Using Ensembles of Multi-Resolution EfficientNets with Meta Data*. <http://arxiv.org/abs/1910.03910>
6. He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. <http://arxiv.org/abs/1512.03385>
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (n.d.). *ImageNet Classification with Deep Convolutional Neural Networks*. <http://code.google.com/p/cuda-convnet/>
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. <http://arxiv.org/abs/2103.14030>
9. Perez, F., Vasconcelos, C., Avila, S., & Valle, E. (2018). *Data Augmentation for Skin Lesion Analysis*. <https://doi.org/10.1007/978-3-030-01201-4_33>
10. Rawlinson, D., Rajagopalan, C., Goldberg, E., & Kowadlo, G. (n.d.). *Deep learning in a bilateral brain with hemispheric specialization*. <https://doi.org/10.48550/arXiv.2209.06862>
11. Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. <http://arxiv.org/abs/1409.1556>
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). *Going Deeper with Convolutions*. <http://arxiv.org/abs/1409.4842>
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the Inception Architecture for Computer Vision*. <http://arxiv.org/abs/1512.00567>
14. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). *Training data-efficient image transformers & distillation through attention*. <http://arxiv.org/abs/2012.12877>
15. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). *PVT v2: Improved Baselines with Pyramid Vision Transformer*. <https://doi.org/10.1007/s41095-022-0274-8>
16. Yang, J. (2023). Leveraging CNN and Vision Transformer with Transfer Learning to Diagnose Pigmented Skin Lesions. In *Highlights in Science, Engineering and Technology CMLAI* (Vol. 2023).
17. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E., Feng, J., & Yan, S. (2021). *Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet*. <http://arxiv.org/abs/2101.11986>
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. <http://arxiv.org/abs/2010.11929>