An Interpretable Transformer Based Approach to Classify Malaria From Blood Cell Images

by

Mehafuza Islam
17301228
S.M. Abdulla Al Mamun
17301031

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 and Engineering

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

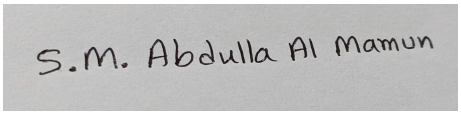
Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing 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:

Mehafuza Islam
17301228



S.M. Abdulla Al Mamun
17301031

Approval

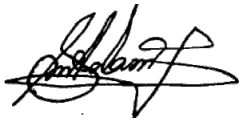
The thesis titled "An Interpretable Transformer Based Approach to Classify Malaria From Blood Cell Images" submitted by

1. Mehafuza Islam(17301228)
2. S.M. Abdulla Al Mamun(17301031)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 17, 2023.

Examining Committee:

Supervisor:
(Member)



Dr.Md.Ashraful Alam
Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul, PhD
Associate 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

Malaria is a disease that can be fatal, and it is spread through the bite of the female Anopheles mosquito. The life of the sufferer is put in jeopardy as a result of the presence of numerous plasmodium parasites, which spread throughout their blood cells. Malaria can potentially be fatal if it is not treated within the first few stages of the disease. A well-known method for diagnosing malaria, microscopy involves taking blood samples from the patient, calculating the number of parasites, and counting the victim's red blood cells. Nevertheless, the procedure of microscopy takes a lot of time, and, in certain circumstances, it can produce an incorrect result. When compared to the more conventional approach of microscopic examination, the recent successes of deep learning (DL) in the field of medical diagnosis make it quite conceivable to reduce the expenses associated with the diagnosis while simultaneously improving overall detection accuracy. This study proposes a transformer-based DL technique for diagnosing the malaria parasite using blood cell images. An explainable AI technique called Grad-CAM was applied in order to determine which aspects of an image the proposed model paid significantly more attention to in comparison to the other aspects of the image through saliency mapping. This was done in order to demonstrate the usefulness of the models. According to the findings of this research, the performance of the vision transformer and the vgg-16 are identical. Both models have reached an accuracy score of approximately 96%, which is very impressive.

Keywords: Malaria parasites; Vision transformer; Deep learning; Grad-CAM

Acknowledgement

We would like to express our heartfelt gratitude to our esteemed thesis supervisor Dr.Md.Ashraful Alam Sir,Our Research Assistant Shakib Mahmud Dipto and CVIS Lab who have consistently supported and guided us through a difficult issue. We were able to overcome all barriers and challenges through their excellent recommendations and regular feedback. Despite the ongoing pandemic, they always made time for us, often at incredibly late hours, and we shall be eternally grateful for the kindness given to us.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ϵ Epsilon

AUC Area Under the ROC Curve)

BF Bilateral Filtering

CAD Computer Aided Diagnosis

CCT Compact Convolutional Transformer

CNN Convolutional Neural Network

CNN Convolutional Neural Network

DBN Deep Belief Network

$GELU$ Gaussian Error Linear Unit

$GRAD - CAM$ Gradient-weighted Class Activation Mapping

LN Layer Norm

Mps Malaria Parasites

MSA Multi-head self attention

NLP Natural language processing

RBC Red Blood Cell

ROC Receiver Operating Characteristic Curve)

SDG Sustainable Development Goal)

$SOTA$ State of Art

SVM Support Vector Machine

TL Transfer Learning

VGG Visual Geometry Group

ViT Vision Transformer

VQA Visual Question Answering

WBC White Blood Cell

Chapter 1

Introduction

1.1 Thoughts behind the Model

A female Anopheles mosquito's bite is how malaria is spread. This awful condition might be lethal. Many Plasmodium parasites procreate within the victim's blood cells, where they survive in a fragile state. Malaria has the potential to be fatal if treatment is delayed when the illness is still in its early stages. The majority of the computerized systems that have been made to this point used traditional deep learning or classic machine learning techniques, which produced serviceable results but still have room for growth. Since the creation of the vision transformer model, the attention-based transformer model has shown promising results in medical imaging, bioinformatics, and other computer vision applications, in contrast to the conventional convolution-based deep learning model. These outcomes were attained following the development of the vision transformer idea. To yet, no attention-based research has been conducted to identify malaria parasites, though. The interpretability of the deep CNN model is once more a significant problem [1]. Recently, the community has become interested in the challenge of graphically expressing the data that a deep learning model has learnt. The interpretability of the model for malaria parasite identification, however, is not included in the vast majority of earlier investigations. In this area, work needs to be done. The paradigm presented in this work uses explainable transformers to distinguish the malaria parasite from blood smear images of cells. This is done in an effort to avoid the issues that have been brought up. Various hyperparameters were assessed in research to enhance performance. They also contained learning rate, weight decay, batch size, and other hyperparameters, in addition to optimizer settings.

1.2 Aims and objective

The aim is to identify the malaria parasite from blood cell images using Vision Transformer,Swin Transformer[1] and VGG-16 [2] models.The gradient-weighted class activation map (Grad-CAM) [3] technique was utilized in order to generate a heatmap image. This image demonstrated the regions of an image to which the suggested model paid significantly more attention compared to the other regions of the image. This was done in order to demonstrate how successful the model that was proposed is. The primary goal is to get the best possible outcomes utilizing the proposed method, which, once various hyperparameters were optimized, should be able to

outperform the state-of-the-art (SOTA) approaches for the identification of malaria parasites.

1.3 Research Objective

Plasmodium, the parasite that causes malaria, enters the human body by bites from female anopheles mosquitoes. It is then disseminated to other individuals by mosquitoes that bite malaria patients, despite the fact that the disease cannot be passed from one individual to another. In addition to the transmission of the disease from mother to child during pregnancy, patients can get malaria via receiving blood transfusions or sharing syringes. A person who is infected with the virus may experience symptoms similar to those of the flu, as well as a high fever, chills, septicemia, pneumonia, gastritis, enteritis, nausea, vomiting, and even death. This is in addition to the flu-like symptoms. Anopheles mosquitoes, which are responsible for the transmission of deadly diseases, prefer an environment that is hot and humid and is located in close proximity to natural water sources. Malaria is most commonly found in areas that fit this description. Because of this issue, the primary objective of this project is to develop a computerized malaria detection framework. This framework will be able to detect parasites in blood smear [4] photographs by replicating the conventional diagnosis of the disease, which is considered the gold standard. Malaria specialists can benefit tremendously from the utilization of such a diagnostic procedure for the aim of making a diagnosis. An automated diagnosis system has the potential to significantly boost the productivity of pathologists and reduce the necessity for highly skilled pathologists in geographically isolated places.The well-known dye known as Giemsa is used to stain peripheral blood smear photographs in order to facilitate the successful visual detection of malaria parasites[5] during the process of microscopy diagnosis. The technique of staining gives healthy RBCs a subdued hue, while giving parasites, platelets, artifacts, and white blood cells (WBCs) a vibrant purple coloration. In order to diagnose malaria, it is necessary to differentiate between blood cells that have been infected by parasites and other components of blood cells[6] that are not infected by parasites, such as platelets, artifacts, and WBCs. This must be done using visual cues. As a consequence of this, the topic of this research is a system based on deep learning that is capable of providing computer-assisted malaria diagnosis. According to estimates provided by the World Health Organization (WHO), this parasite is responsible for between 300 million and 500 million cases of infection each year, affecting more than two million individuals. Red blood cells, often known as RBCs, are the major focus of the malaria parasite that lives in human blood. Under a light microscope, a pathologist examines the red blood cells (RBCs) [7] in order to examine the changes in their size, shape, and color. This is done in order to diagnose malaria. The level of accuracy with which a microscopy diagnosis of malaria can be made will depend on the pathologist's familiarity with the disease. As a result of the subjectivity involved, this method is not only inefficient but also wrong, all of which can lead to a diagnosis that is inconclusive and inaccurate, an inappropriate course of treatment, and possibly even the patient's death. The rapid administration of the appropriate treatment is made possible by the early diagnosis of malaria infections by medical personnel. Because of this, we have decided to concentrate our efforts on the procedure that will be described below.

- Various transformer based models were employed for the very first time to detect malaria parasites.
- The gradient-weighted class activation map (Grad-CAM) method was utilized in order to interpret and visualize the trained model.
- Malaria parasite datasets were used for experimental analysis.
- SOTA models and the suggested model for parasite detection in malaria were contrasted

Chapter 2

Background Study

2.1 Literature Review

Malaria is a disease that can be fatal, and it is transmitted by the bite of the female anopheles mosquito.The life of the sufferer is continually put in peril as a result of the proliferation of various plasmodium parasites in their blood cells. In the absence of treatment in its early stages, malaria can be lethal. According to the World Health Organization, malaria was responsible for the deaths of about 438,000 and 620,000 people in 2015 and 2017, respectively, with an additional 300 million to 500 million people being affected [8]. The first method is laborious and time-consuming since it involves the identification of at least 5,000 RBCs, while the second is an expensive antigen-based fast diagnostic screening. Both of these procedures are traditionally used to diagnose malaria. The diagnosis of malaria, the collection of blood samples from patients, and the tally of parasites and red blood cells are all typically accomplished by the use of microscopy. However, the method for microscopy takes a lot of time and can provide inaccurate results in some circumstances. In recent years, academics have focused their efforts on finding a solution to this problem by applying a variety of ML and DL methods[9]. This is to circumvent the limitations that are imposed by traditional methodologies. It is now possible, thanks to the recent success of machine learning and deep learning in medical diagnosis, to reduce diagnostic costs while improving overall detection accuracy using a multi headed attention-based transformer model when compared to traditional microscopy methods [10]. This is made possible by the fact that it is now possible to use machine learning and deep learning. The gradient-weighted class activation map (Grad-CAM) technique can be used to generate a heatmap image that identifies which regions of an image the suggested model paid much more attention to than the remaining parts. This can be used to demonstrate that the model was successful. A sizeable number of computer vision studies for automatic parasite identification of malaria utilizing images of peripheral blood smears have been presented in the relevant research literature. [11] presented a method for the recognition of stained objects that was based on a color histogram. This was in addition to providing a suggestion for a binary classification approach that was based on the K-nearest neighbor algorithm. Later, a method for classifying peripheral blood smears that included Plasmodium parasites was developed and used [5]. A novel thresholding approach that is founded on a morphological technique was proposed in [12] as a means of determining whether or not an individual is infected with

malaria parasites. [4] made the suggestion that segmenting parasites use the HSV color scheme. The malaria parasites were located using a statistical technique [12]. [13]] proposed using mathematical morphology and granulometry to detect parasite infestations. The accuracy of the diagnosis would increase as a result. It has been demonstrated that large-scale NLP models significantly improve language task performance without displaying any signs of saturation. They are equally adept at firing one to four shots, much like humans. The objective of this research is to investigate massive models in the field of computer vision. Three of the most important problems that come up during the training and deployment of large vision models are the need for labeled data, resolution gaps between pre-training and fine-tuning, and training instability [14]. Three fundamental tactics have been offered, and they are as follows: The first method efficiently transfers models that have been trained on low-resolution images to tasks that later need high-resolution inputs by combining the residual-post-norm approach and cosine attention. The second approach uses a log-spaced continuous position bias mechanism. The third method, known as SimMIM, eliminates the need for big annotated images by using a self-supervised pre training process. Before a large-scale and generalized vision model can be successfully trained, a few important issues must be fixed. Our examination of very big vision models has shown that there is a difficulty, which starts with the stability of the training. When dealing with larger models, the differences in activation amplitudes between layers can occasionally be seen clearly [15]. Because the output of the residual unit was immediately added back to the main branch, this was the result. This was discovered after carefully examining the original architecture. When layers are piled on top of one another, activation values rise and the amplitudes of later levels are plainly greater than those of early layers. [16] suggested the res-post-norm normalization alternative as a remedy for this problem. In this case, each residual unit's LN layer is moved from the front to the back. High quality input images and attention windows are crucial for many downstream vision tasks, including object detection and semantic segmentation. There may be large changes in window size between low-resolution pre-training and high-resolution fine-tuning [17]. The bicubic interpolation method is currently used to create the location bias maps. This quick repair usually has negative results because it wasn't well thought out. Because it supports any coordinates, a pre-trained model can instantly switch between window sizes by sharing the weights of the meta network. This is represented by the log-spaced continuous position bias, or Log-CPB. There is presently no segmentation dataset for RBC cell images with a parasite mask for use in quantitative research. The dataset used in this work was annotated to show that photographs of typical RBCs had a version that was clean and error-free, in contrast to images of parasites. Depending on whether these flaws were present or not, the images of RBC cells were labeled as parasitic or normal. Heatmap pictures were created using the GRAD-CAM [3] method to show how easily the trained model might be described. These visualizations displayed the specific areas of the image that the model concentrated on while doing feature extraction and classification. This approach might provide new insights that facilitate precisely identifying MPs.The RBC images were from an open-source source. But in order for the procedure to work fundamentally, RBCs must first be distinguished from other cells in blood smear images [3]. The RBC regions that the malaria parasite has damaged are then used to classify the parasite. In this study, segmentation was skipped in favor of using photographs of generated

RBCs to categorize the malaria parasite. Conventional CNN models have an inductive bias that is exclusive to some images because the local receptive field theory was employed to create them. CNN models require exceptionally deep network models or larger kernels to gather data on a global scale. The given model for the malaria parasite, which was based on a transformer, performed superbly since transformer models are not restricted by these restrictions. The Grad-CAM image additionally demonstrated how straightforward its dispersion was[18]. In fine-grained identification tasks including object recognition at the region level, semantic segmentation at the pixel level, and image classification at the picture level, a general-purpose computer vision backbone known as the Swin Trans- former[1] ,[19] shines. Swin Transformer's major objective is to combine the advantages of both systems by adding a number of significant visual priors, including as translation invariance, hierarchy, and locality, to the original Transformer encoder [11]. Strong modeling capabilities and a user-friendly interface for several visual operations are additional features of the basic Transformer unit [8]. We combined Grad-CAM with the currently in use fine-grained visualizations to create a high-resolution class-discriminative visualization. The models for captioning, visual question answering (VQA), and image categorization are then implemented using this format. Our graphics outperform existing approaches in the weakly-supervised localization task ILSVRC-15. By highlighting dataset bias, they also improve representations of the underlying model and encourage model generalization [20].. They also draw attention to the flaws in these models by demonstrating that predictions that appear illogical may actually have rational causes. Furthermore, our representations perform better on the ILSVRC-15 weakly-supervised localization issue than earlier techniques. Our visual representations of photo captioning and VQA demonstrate that even algorithms without an attention component may localize inputs. To determine if Grad-CAM [21] explanations assist users in placing the proper degree of trust in deep network forecasts, we construct and conduct human testing. Our findings show that Grad-CAM can assist novices in correctly identifying a deep network prediction. The promise of Grad-localization CAM in the context of picture classification. Competing systems must offer bounding boxes and classification labels in order to meet the ImageNet localization challenge[22]. Evaluation is carried out for the top-1 and top-5 categories in the anticipated list, much like categorization. From a picture of [3], the class's first assumptions are made. Grad-CAM maps are then produced for each of the predicted classes using a threshold of 15% of the maximum intensity after the data has been binaryized. As a result, connected pixel segments are produced. The bounding box is established using the section that is the largest on its own. Intuitively, a good explanation for a prediction is one that strengthens the discriminative representations of the targeted class. The experiment produced 360 unique graphics by rendering each of the 90 image-category pairs in one of the four potential ways. By first collecting ratings for each image, comparing those evaluations to the actual data, and then averaging the results, the accuracy was evaluated. Compared to Guided Backpropagation, which helps participants identify[22]. the category in just 44.44% of cases, Guided Grad-CAM helps participants identify the category in 61.23% of cases. Therefore, guided grad-CAM enhances human performance by 16.79%. Grad-CAM increases the class discrimination of Deconvolution from 53.33 to 61.23 percent. Guided Grad-CAM outperforms all alternatives by a significant margin. Importantly, our findings tend to show that Deconvolution is superior[23]

to Guided Backpropagation at differentiating between multiple classes, despite the fact that it has a more visually appealing appearance. In our opinion, our research is the first to quantitatively discover this exceedingly little variation. The objective of this multi-head attention-based transformer model is to detect malaria parasites. Grad-CAM visualization, which offers an interpretation of the learnt model, was also used to validate the learning [17].

2.2 Related Works

Recent studies have concentrated on employing photo analysis powered by artificial intelligence to detect malaria (AI). Malaria parasites (MPs) can be located in RBC pictures using a deep belief network (DBN)[20]. After being trained on 4100 images, their model performed remarkably well, obtaining an F-score of 89.66%, a specificity of 95.92%, and a sensitivity of 97.60%. An artificial neural network is utilized to identify MPs from photos of RBCs. They were able to train their algorithm to reach an accuracy of 90% to 100% using pictures of healthy and sick RBCs. We successfully identified MPs from RBC images with a 97% accuracy rate and a low-cost identification technique that eliminated the use of GPUs or any preprocessing techniques [7]. CNN models achieved an accuracy of 92.7% when it comes to extracting attributes from 27,558 RBC cell images and identifying MPs [22]. A simple CNN was developed using RBC pictures, and it was effective at differentiating MPs [5]. Using 19,290 training images and 4134 test datasets, the model was trained to an accuracy rate of 98.85%. The MPs were identified using the Xception, Inception-V3, ResNet-50, NasNetMobile, VGG-16, and AlexNet transfer learning models (TL) [8]. When several optimizers and activation functions were merged, the model performed better. They trained their models using 7000 different pictures, and the overall accuracy of their models is 99.28%.It was suggested to use a fake CNN model to locate MPs after establishing which RBC photos had been incorrectly classified [11]. To increase the model's accuracy, three training methods—general, dis- tillation, and auto encoder training—were applied. [15] suggested using a mobile application and a cyclical stochastic gradient descent optimizer to discover MPs. They discovered that the model's accuracy rate was 97.30%. A unique CNN model for the identification of MPs was developed by integrating bilateral filtering (BF) and photo enhancement techniques [4]. Accuracy was 96.82% overall. Using thin RBC images, a layered CNN model was developed to predict MPs.
They performed at a phenomenal level, with accuracy, precision, and recall of 99.98%, 100%, and 100%, respectively. Hung and Carpenter suggested employing a CNN that was especially tailored to understand what was displayed in the RBCs. One-stage classifications had an overall accuracy of 59%, whereas two-stage classifications had an accuracy of 98%. A system that could diagnose malaria using images of the wounded cells was developed utilizing computer assisted diagnostics (CAD), also known as CAD[10]. By pretraining the parameters of an artificial neural network with a functional connection and sparse stacking, they were able to recognize malaria from a private dataset of 2565 RCB pictures collected from the University of Alabama at Birmingham with an accuracy of 89.10% and a sensitivity of 93.90%. The University of Alabama in Birmingham provided the dataset. Using a CNN and an SVM, it is possible to get accuracy ratings of 91.66% and 95%, respectively. [15] asserts that a variety of photos were used to modify CNN. The cell counting problem

was converted into a segmentation problem, and a two-level segmentation solution was produced. The output of the CNN focus stack model has 98.77% accuracy, sensitivity, and specificity rates[12]. To predict MPs from RBC pictures, three distinct machine learning (ML) models have been developed: logistic regression (LR), decision tree (DT), and random forest (RF) [5]. They first employed RF to extract the composite features from the cell photos in order to achieve a high recall of 86%. A computer-aided detection system (CAD) was created to find MPs in RBC images after the background noise was reduced and the overall image quality was improved using the BF approach [10]. They used adaptive thresholding and morphological image processing to detect MP with a 91% accuracy rate. Self-organizing maps (SOM) and adversarial evolution (AE) were both utilized in the search for MPs, however it was shown that AE was more accurate than SOM, which had an accuracy of just 87.5%. LeNet, AlexNet, and GoogLeNet were the three TL models made available by for the purpose of identifying MPs. The accuracy of the support vector machine was 92%, whereas that of the TL models was 95%. The two were compared using the SVM (SVM). Grayscale preprocessing was performed to increase contrast and global thresholding to remove various blood cell components from the photos in order to distinguish MPs from RBC images. These two techniques can be used to locate MPs in RBC pictures. A CNN model with a 96.82% MP detection rate was produced using a combination of bilateral filtering (BF) and picture augmentation techniques. As a result, we were able to create a CNN model that met our requirements. It is necessary to create a layered CNN model in order to predict MPs from thin RBC images. With a recall of 99.9%, an accuracy of 99.98%, and a precision of 100%, they performed approximately averagely. The object in the RBC pictures was recognized using a CNN constructed on geographic locations. With an accuracy of 89.10% and a sensitivity of 93.90%, computer-aided diagnosis (CAD) [8] was used to identify malaria from a private dataset of 2565 RCB images acquired [24]. Overall classification accuracy for one-stage classifications was 59%, while accuracy for two-stage classifications was 98. A approach for identifying malaria from a cell picture uses an artificial neural network with a functional connection and sparse stacking. Accuracy values of 95% and 91.66%, respectively, using a mixed support vector machine (SVM) and convolutional neural network, were obtained (CNN).

2.3 Used Architectures

2.3.1 Vision Transformer

Vision Transformer[11] is the name given to the method that places an emphasis on the classification of images and employs a design that resembles a transformer across the different areas of a picture. After being broken up, the pieces of a fixed size that make up an image are then linearly embedded. Following the incorporation of positional embeddings, the completed vector sequence is fed into a conventional transformer encoder. Vision transfer has recently emerged as a viable alternative to the more traditional convolutional neural network (CNN)[22]. CNN was commonly used for various image classification tasks until ViT models exceeded CNN by around 4 times in terms of the accuracy and efficiency of computation. Vision Transformer was used to perform image classification, object recognition, and semantic image segmentation in recent benchmarking tests for a variety of computer vision systems.

These tests were conducted to evaluate the performance of various computer vision systems. When compared to CNNs, we find that Vision Transformer has a significant reduction in the amount of image-specific inductive bias. In CNNs, each layer of the entire model is baked with locality, two-dimensional neighborhood structure, and translation equivariance. This is done so that the model can perform optimally. Only the MLP levels are local and translationally equivariant in ViT; the self-attention layers are global throughout the rest of the architecture. During the process of fine-tuning the model, the position embeddings are altered to take into account images of varying resolutions. On the other hand, the two-dimensional neighborhood structure is utilized only very occasionally. Aside from that, all spatial interactions between the patches need to be learned from the beginning because the position embeddings at the time of initialization carry no information about the patches' 2D positions. Transformers are incapable of working with grid-structured data because they are unable to perceive any spatial relationship between the characteristics of the data. Instead, in order for them to function, they are required to operate utilizing sequences obtained from a non-sequential spatial signal. First, the image is fed into the transformer, which cuts it up into predefined patches that are exactly the same as words or sequences and are thus encoded using the same word encoder. The supplied image is then turned into an N-dimensional patchwork pattern. Additionally, N might stand for the length of the sequence or the number of words. In addition to this, it employs a constant hidden vector of size D across the layer in order to circumvent the issue of vanishing gradients by making use of skip connections. On the other hand, there is no decoder; rather than that, a specific MLP is used in order to produce the final output. The result of applying such patches is represented by the patch embeddings. The second step that they took was to add a learnable embedding on top of the output sequence. This learnable embedding has a state of its own, and at the output of the transformer encoder[8], it serves as the representation of the input image. Adding a classification head to the transformer encoder is done both during the pre-training phase and the fine-tuning phase. The MLP that serves as the classification head employs a single layer for fine tuning and a hidden layer for use during the pre-training phase. This is exactly the same as the class token for the BERT. To ensure that the positional information associated with the patches is not lost, standard 1D positional embeddings are appended to the patch embeddings. The next component of the network is called the transformer encoder. It is comprised of MLP blocks, each of which consists of two layers, a GELU non-linear, and alternating layers of multiheaded self-attention (MSA). To add LN, Layernorm, and residual connections before and after each block, the following equations are utilized. It appears that one of the greatest benefits of vision converters is
MLP blocks only contain the local and transitionally equivariance parts, whereas CNNs have them (the transitional equivariance) in each layer throughout the overall network alongside 2d neighborhood structure. This means that MLP blocks are nearly free of inductive bias associated with specific images. Only the first step, which is when the images are divided into patches, and the fine-tuning phase, which is when the positional embeddings are updated, incorporate locality or the 2d neighborhood structure. The layer of self-attention, which is the most important component, is global in nature. Last but not least, the network is flexible in the sense that, rather of feeding it raw images, we can provide it CNN output such as feature maps. The patch embedding projection E has been linked to the output of

the CNN feature map that was fed into it. When trained on more extensive datasets, the vision transformer is able to attain dramatically improved performance. After training, it needs to have some fine-tuning done in order to do classification tasks in which the D-K feedforward layer is used instead of the prediction head (K is the number of classes). When compared to pre-training, fine-tuning the transformer using images of a greater resolution provides extra benefits. This is due to the fact that detailed photos produce a higher word count, which is a direct result of the constant patch size. The inductive bias problem is not manually introduced until the patch extraction and resolution adjustment phases, respectively.
The basic element of the compact convolutional transformer is the multiheaded self-attention, which acts as primary component (MSA). It is not essential to examine the complete portion of an image in order to extract information that is valuable; rather, the attention mechanism focuses on the area of the image that is useful. There have been numerous attentional mechanisms developed up to this point. Having said that, the multiheaded form of self-attention was initially presented in the context of the vision transformer.
The scaled dot-product attention function can be expressed in the following shorter form:

$$Attention(Q, K, V) = softmax((Qk^T/\sqrt{d_k})V \tag{2.1}$$

$$MultiHead(Q, K, V) = Concatenate[head_1, head_2, ...head_h]B^0 \tag{2.2}$$

Where Q,K,V are queries, keys, and valuesrespectively.
After the picture patches have been processed by the convolutional block, they are sent on to the transformer encoder for further processing. The application of layer normalization to the picture patches is the first step that must be taken in order to normalize the activations along the feature direction as opposed to the mini-batch direction, which is the direction that is used in batch normalization. The subsequent step is to provide these areas of normalization with multiheaded self-attention. The results are appended, along with the remaining original patches that are relevant to the investigation. An additional residual connection, additional layer normalization, and a feed-forward block are all mandated by this regulation. The feed-forward block is made up of several different components, including a dropout layer, a GELU nonlinearity, two linear layers, and others. The dimension is increased by a factor of four in the first linear layer, but it is decreased in the second linear layer (feed-forward block).

$$M' = Linear(C\epsilon R^{N\times d})\epsilon R^{N\times 4d} \tag{2.3}$$

$$M'' = Dropout(GELU(M'))\epsilon R^{N\times 4d} \tag{2.4}$$

$$M' = Linear(M''\epsilon R^{N\times 4d})\epsilon R^{N\times d} \tag{2.5}$$

These two pathways' results are combined once more. 8 transformer encoders were successively assumed below:

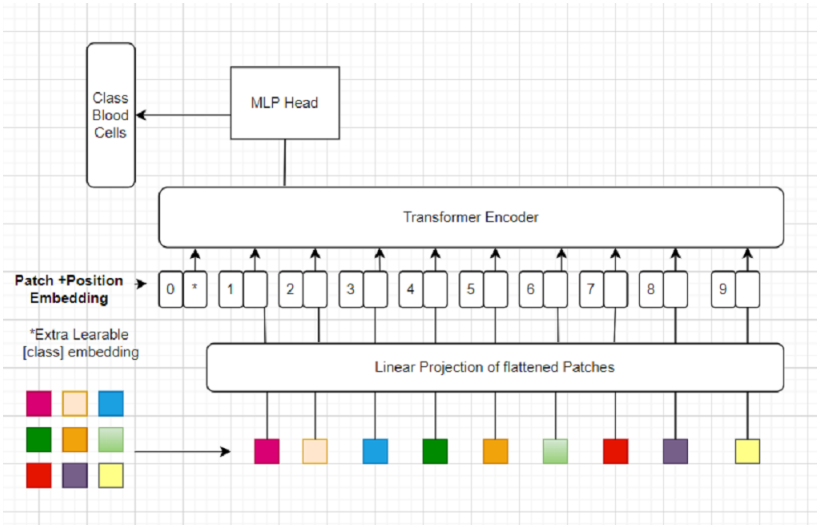


Figure 2.1: Vision Transformer(Transformer Encoder)

2.3.2 Swin Tranformer

A subtype of the vision transformer known as the swin transformer has windows that are moved in different directions [1]. An input RGB image is split into non-overlapping patches by a patch-splitting module that is quite similar to vision transformer. This is the first step in the Swin Transformer process. Tokens are the name given to these fragmented patches, and each token has its unique characteristics that are determined by concatenating the raw RGB values of the pixel it belongs to. Due to the fact that the size of the patch is always a constant 4 by 4, the feature dimension of each of these tokens is the number 48. Following that, these features are projected to an arbitrary dimension C utilizing linear embedding layer in the same manner as the initial vision transformer. In addition, several Transformer blocks that have modified self-attention computation are utilized on these patch tokens [19]. These Transformer blocks are referred as as Swin Transformer blocks. They preserve the 11 token count as the result of the HW/4, which also completes stage 1 of the transformer alongside the linear embedding. As the depth of the network increases, the number of tokens is reduced through the process of patch-merging layers in order to create a hierarchical representation. The leading patch merging layer will concatenate the features[25] of each set of two adjacent two-by-two patches as it does its work. In addition to this, a linear layer is applied to each of the 4C-dimensional features. As a result of this, the total number of tokens has been cut down by a factor of two times two, which is equal to four. This results in a resolution reduction of two. The output dimension will always be 2C, regardless of whether the resolution is changed from H/8 $\times$W/8 or not. The second stage of the transformer is comprised of the initial patch merging as well as the feature transformation. Now, in order to gather more windows, the process described above is carried out again, first with an output resolution of H/16 $\times$W/16 and then again with an output resolution of H/32 $\times$W/32. Because these repeated phases work together to provide a hierarchical representation with the same feature map resolution as the standard convolution neural networks like the VGGnet[2] or Resnet that have been discussed[26], they are able to easily replace the networks that are currently being used for a variety of works relating to computer vision.

Here we are at the brand new swin transformer block that was just unveiled. Each one of them is built with a shifted window-based MSA module, which is then followed by a double-layered MLP that includes a GELU nonlinearity between them and one another. In the same manner as with VIT, each MSA module and MLP is preceded by a LayerNorm (LN) layer, and each module is followed by a residual connection in order to resolve the issue of gradient vanishment. The next step is called Window Based Self-Attention (W-MSA), which comes after that. Calculation of the self-attention is carried out when the window is in this local scope. After that, the Shifted Window Partitioning in Successive Blocks (SW-MSA) algorithm will be broken down for you.

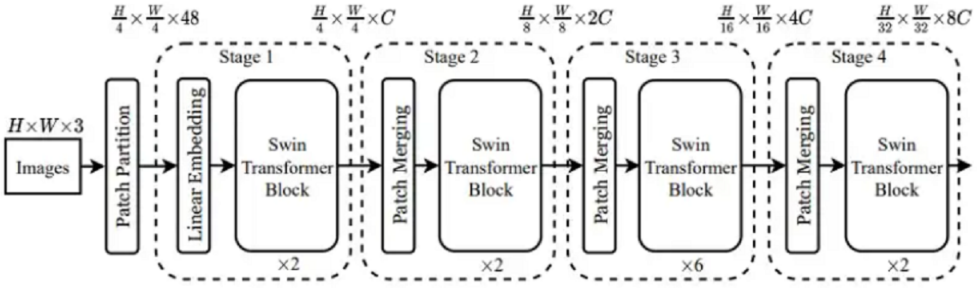


Figure 2.2: Swin Tranformer Model Architecture

This window-based self-attention module does not include support for cross-window connections, which, as a result, reduces the module's overall modeling capabilities [1]. A shifted window partitioning technique is recommended for use in consecutive Swin trans-former blocks. This strategy switches back and forth between the two configurations of the partitioning window. Although a traditional method of window partitioning is utilized in the first module, the succeeding module utilizes a windowing configuration that is distinct from that of the layer that came before it. This is accomplished by moving the windows to new locations. After that, the successive swin transformer blocks are computed utilizing the output of the layer that came before it. This movement of partition windows would result in the creation of "new" windows, which would make the entire network more complicated. One strategy that is utilized to address problems of this nature is called cyclic shifting towards a goal. A more effective method for doing batch calculation is presented here in the form of cyclic shifting in the direction of the top-left corner, as was shown previously. This cyclic technique creates the same number of batched windows as the regular window partitioning, which allows it to avoid performance concerns, which in turn contributes to the method's efficient operation. There is a selection of numerous models to choose from, each one matched to the complexity and necessary hardware. The Swin-B, which is the foundational model, comes in a standard model size and has a level of complexity comparable to that of a conventional vision transformer. The T suffix model is the smallest one, with a quarter of the time complexity; the S suffix model is the small one with half; and finally, the L suffixed model is the largest one, with double the complexity of the base model and requiring similar capacity to deploy ResNet-101. Using a patch splitting module, such as ViT by Swin Transformer[25], an RGB input image is split into non-overlapping patches. Each patch has a "token" feature that is set up to represent the sum of the RGB values of all the individual pixels. Each patch in our implementation has

a feature dimension of 4 by 4 by 3 or 48 because we utilize a patch size of 4 by 4. A layer of linear embedding enables the projection of this raw-valued feature to any dimension. These patch tokens make extensive use of Swin Transformer blocks, which are Transformers with better self-attention computation. As the network depth increases, a hierarchical representation can be created by lowering the overall token usage during the patch merging process. The first patch merging layer adds a linear layer to the concatenated features of the 4C dimension after concatenating the features of each set of two adjacent patches. Because 4 is a multiple of 22, the resolution is scaled down by a symbolic factor of 4. An output dimension of 2C was chosen. The multi-head self attention (MSA) module that is typically featured in a Transformer block is replaced with a module based on shifted windows to create the Swin Transformer. The remaining Transformer blocks share the same layers. A Swin Transformer block is made up of the GELU nonlinearity, a shifted window-based MSA module, and a 2-layer MLP. Each MLP is followed by a residual connection, and each MSA module is preceded by a LayerNorm (LN) layer. Additionally, a connection is added after every module. Then, we give each head a relative position bias of B: First, we compute self-attention in line with [4].

$$Attention(Q, K, V) = softmax(Qk^T/\sqrt{d} + B)V \tag{2.6}$$

2.3.3 VGG-16

VGG-16, a variation of the CNN (Convolutional Neural Network) model, is currently considered to be one of the best computer vision models available [2]. The creators of this model conducted an analysis of the networks and improved the depth by utilizing an architecture with exceptionally tiny (3×3) convolution filters. This model's results indicated a significant development over the state-of-the-art setups. After increasing the depth to between 16 and 19 weight layers, there were approximately one hundred and thirty Eight trainable parameters obtained. The object identification and classification approach known as VGG-16 has a success rate of 92.7% when applied to the task of categorizing 1000 photographs into 1000 distinct groups [11]. A deep convolutional neural network with 16 layers, which is regularly fused with 33 convolutional layers and 22 pooling layers, is one of the image classification methods that is utilized extensively. It is a deep neural network with 16 layers. The ability of VGG-16 to effectively extract features from images is what makes it possible for it to do well in image classification. During our test, the VGG-16 obtained an accuracy rate of 95%.

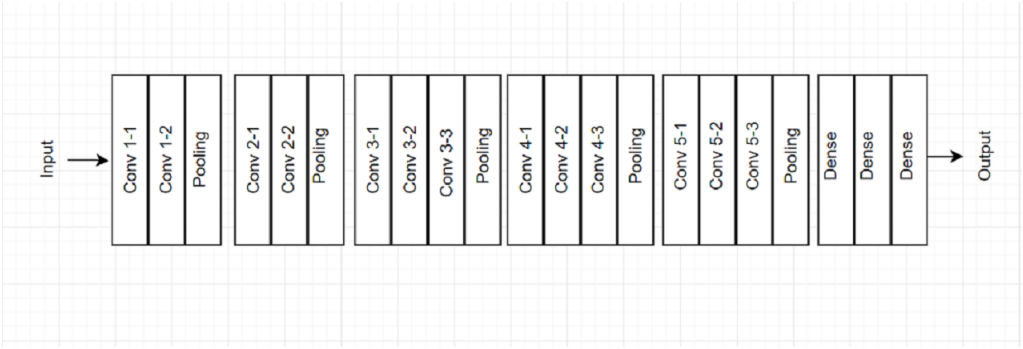


Figure 2.3: VGG-16 Model Architecture

VGG16 is a variant of CNN (Convolutional Neural network) that was introduced by two members from Oxford University's Visual Geometry Group Lab in response to the yearly ILSVRC (ImageNet Large Scale Visual Recognition Challenge) challenge. As a result of the model's efficiency and ground-breaking performance, it was awarded both the first and second prize for the competition. VGG16 was developed by two members from Oxford University's Visual Geometry Group Lab. The numerical suffix indicates the number of layers, which in this case is 16, that have parameters that the model is able to learn from. It does not have a big number of hyperparameters like concurrent networks have, but it does choose convolution layers with a size of 3 3 (stride = 1) and padding and max-pooling layers with a size of 2 2 (stride =2) respectively. VGG16 is a method for identifying objects and classifying them that has a rate of accuracy of 92.7% when classifying a thousand photographs into a thousand separate categories, which totals over fourteen million images. It has quickly become one of the most popular image classification methods due to its ease of use, which is made possible by the fact that the pre-trained network comprises more than one million photos[27] taken from the ImageNet database. This model suggested making use of a very compact 3 by 3 receptive field (filters) with a stride of 1 pixel across the entirety of the network. The utilization of 3x3 filters uniformly distributed over the model is a standout characteristic of the network. These filters effectively function as a 5x5 filter when counting two consecutive filters together and as a 7×7 filter when numbering three filters in succession. Although it would seem that the addition of these additional layers would make the model more difficult to understand, the reality is quite the reverse. Additionally, in addition to the three convolution layers, there are also three non-linear activation layers. This is in contrast to the single non-linear activation layer that is present in models that use 7 ×7 filters. This allows the model to discriminate much more effectively and provides the network with the ability to converge relatively quickly. It also brings the total number of weight parameters in the model down by a significant amount. If both the input and output of a three-layer 3 3 convolutional stack contain C channels, then the total number of weight parameters is equal to 3 ×32C2, which is equal to 27C2. When compared to a convolutional layer with 7 7 dimensions, the needed quantity of 72C2 = 49 C2 is approximately double that of a layer with 3 ×3 dimensions. This might also be thought of as a regularization of the 7 ×7 convolutional filters by forcing them to deconstruct through the 3 ×3 filters and providing non-linearity in between through the use of ReLU activations. The tendency of the network to over-fit during the learning phase would be reduced as a result of this change. Numerous network designs based on network depth conducted experiments with a variety of such setups, and the following were among the designs that were entered into the ImageNet Challenge. (typically one, two, or three) convolution layers of filter size 3 ×3, stride one, and padding one, followed by a max-pooling layer with a size of 2 ×2, and finally, a max-pooling layer with a size of 2 ×2. The network parameters were changed to match different configurations of this stack in order to obtain the proper depth levels. A number is assigned to each configuration that indicates how many tiers of weight parameter values are present in that configuration. The two levels before it are each 1,000 pixels wide, whereas the third level is 4,096 pixels. They are all connected. These layers come after convolution stacks. The final layer of the output layer employs soft-max activation. The number 1000 indicates that there are a total of 1,000 classes. The 224 ×224 pixel images in the

collection may all be identified by their red, blue, or green tones. The model's input is this dataset. The resulting tensor has three dimensions and 224 by 224 dimensions. The processed vector, made up of 1,000 values that represent the likelihood of each class and whose cumulative total must equal the total, is then distributed by the network. Following ReLU activations, each image goes through a stack of two convolution layers, the first of which has two exceedingly small receptive sizes of 3 3. Over these two levels, there are a total of 64 filter slots available. The padding and convolution stride can only be as large as one pixel. This is a disadvantage. In this configuration, the spatial resolution is preserved, and the dimensions of the input image are accurately translated into the dimensions of the output activation map. A 2 by 2 pixel frame with a 2 pixel stride is then used to spatially pool the activation patterns. The intensity of the activations is consequently 50% lessened. As a result, the activations at the bottom of the first stack are 112 by 112 times 64. The idea does have some disadvantages, though, such the fact that it is a very inefficient choice because it uses up a lot of bandwidth and storage. Last but not least, the existence of the growing gradient problem is caused by the fact that there are so many parameters—approximately 128 million.

2.3.4 Grad-CAM Visualization

The gradient-weighted class activation map, or Grad-CAM [3], is a technique that can be used to determine what the model has actually learned. This method produces a heatmap that is particular to a particular category by employing a deep learning model that has been trained on a particular input image. This Grad-CAM technique highlights the parts of the input image where the model concentrates on developing discriminative patterns since the final layer comprises the characteristics with the highest amount of semantic significance. The feature mappings produced by the last convolutional layer serve as the foundation for the graduated enhanced CAM's discriminative semantics. Grad-CAM (Gradient-weighted Class Activation Mapping) highlights the crucial regions in the image for idea prediction by exploiting the gradients of each target concept as they flow into the final convolutional layer and producing a coarse localization map. In contrast to earlier methods, GradCAM can be used with a wide range of CNN model-families. For tasks requiring multimodal inputs, such visual question answering or reinforcement learning, as well as for structured outputs, like captioning, this consists of CNNs with fully linked layers (VGG). None of these uses necessitates the use of specialized knowledge or structural changes. Guided Grad-CAM, a high-resolution class-discriminative visualization, is created by combining Grad-CAM with the available fine-grained visualizations. As a result, we can classify our pictures more quickly. We investigate potential causes for the VGG-16 CNN's failure to accurately detect ImageNet images when using Guided Grad-CAM. As a starting point, we are given a collection of samples that the network cannot correctly identify. This allows us to now identify any current network problems. We use a technique called Guided Grad-CAM to graphically depict the right grades as well as the projected classes for the cases that were erroneously categorised.

The Guided Grad-CAM visualization outperforms earlier methods in this research due of its superior resolution and great class discriminating abilities. Data from the training dataset was initially used to train the suggested model. The developed

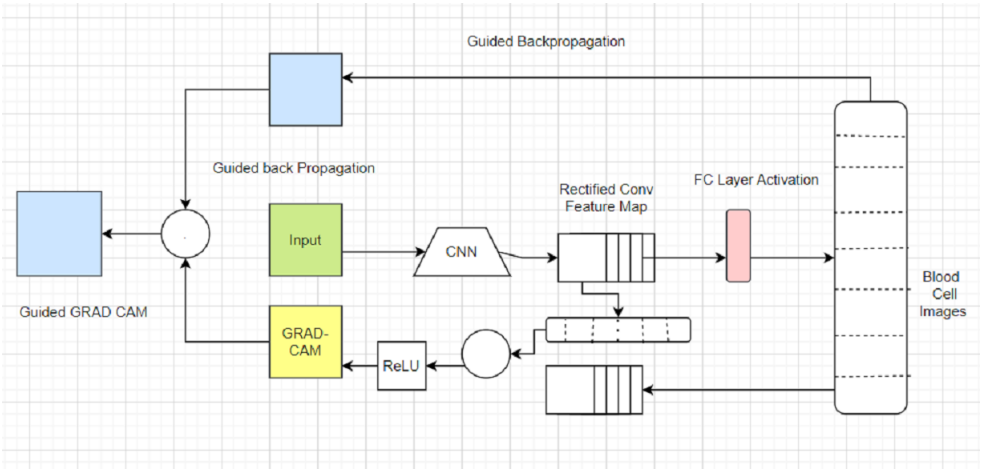


Figure 2.4: GRAD CAM Visualization

model was evaluated using the dataset's testing segments after the training phase. The findings were examined. The actual learning procedure for the trained model using the aforementioned Grad-CAM approach was also presented. The required heatmap was produced using the trained model, a random sample of test pictures, and the Grad-CAM technique. The multilayer perceptron layer of the final transformer encoder, which came before the final classifier, was the target in this instance. After collecting features and gradients from that layer, the heatmap mentioned earlier in this paragraph was made using the Grad-CAM method. A nearest-neighbor interpolation was used to scale the heatmap down until it was the same size as the input image, and it was then placed on top of the original picture. The heatmap was modified using a jet color map. The dots on the lesions are notably more reddish than other parts of the image, as shown in the overlay heatmap photos. These common lesions, which appear as red patches whenever the sickness is present, are brought on by the malaria parasite.

Chapter 3

Research Methodology

3.1 Proposed Methodology

The classification of Malaria based on photographs of blood cells will be the focus of our paper. In order to achieve the level of accuracy we require, we typically train our model using a sizeable portion of the dataset from which it will learn. In the first step of our process, we prepared our dataset. To put it another way, we pre-processed the dataset by removing photographs that were unclear or blurry in order to avoid getting biased results from the suggested model when it was trained using the dataset's training samples. After the training phase had been finished, the model that had been trained was put to use in order to analyze the testing portions of the dataset [9]. In addition, the Grad-CAM technique, which was mentioned before, was applied in order to demonstrate the actual knowledge that the trained model had acquired. In order to build the matching heatmap from the trained model citesselvaraju2017grad, the Grad-CAM method was utilized, and a number of different test images were selected at random. In this particular instance, the multilayer perceptron layer of the very last transformer encoder that came before the final classifier was the layer that was intended to serve as the goal. The features and gradients of that layer were recovered, and a heatmap was generated with the use of the Grad-CAM technique that was discussed earlier. Following that, the heatmap was shrunk to the same size as the input using nearest-neighbor interpolation which is in brief the classification of malaria based on images of blood cells is the objective of our research. To begin, we gathered the dataset, following the completion of the dataset collection,we normalized the pixel values so that they fell within the range 0-1. Then we down sampled the images to a resolution of 64×64 pixels. Next, we performed a 7:2:1 split of the gathered data. A random selection of 70% of the image was used for the training of the models, 20% of the image was used for validation, and the remaining 10% of the image was used for the evaluation phase. We next augmented the images from the training set by randomly flipping them horizontally, zooming in between 20%, and rotating them by 2%. Three deep learning models have been trained, and their results have been compared (VGG16, Vision Transformer, and Swin Transformer). After that, we put Grad-CAM into action in order to analyze the predictions made by that model. The work plan diagram is given below:

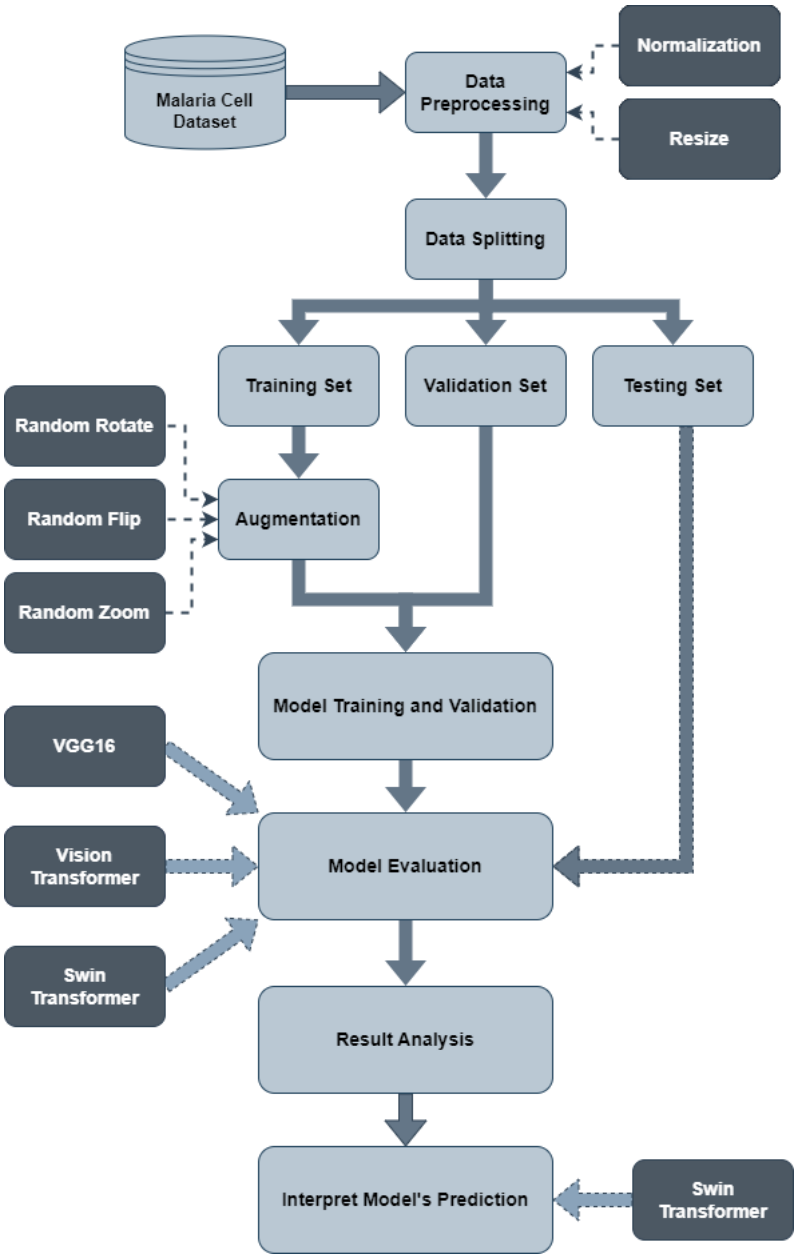


Figure 3.1: Proposed Methodology

i) Pre-processing: When choosing our dataset, we avoided hazy and blurry images and instead chose clear, high-resolution images. We employed the following procedures throughout our data pre-processing to increase the quality and precision of our dataset:
• Scaling: We scaled our raw data size and drove it between 0 and 1 from 0 to 255 in order to fit our dataset into the model.
• Augmentation: In our data, we employed 10 percent of the width shift. The range of floating point numbers utilized for width shift is 0.0 to 0.1.
ii) Splitting: Our dataset has been splitted into three sections. 70percent of the data in the first section is divided for training purposes. The remaining dataset is used to test data, with 20 percent of it being used for validation.
iii) Classification: Following the training dataset, the CNN-based models VGG-16, Vision Transformer, and swin Transformerare used to train and converge the data.
iv) Result: In this case, the test dataset is fed to the model implementing GRAD-CAM and we can then assess how accurate our proposed model is.
Recently, a number of attention-based models have been created. The vision transformer has so far drawn the most attention from researchers for computer vision problems. The vision transformer, first introduced in 2021 [13], has been slightly changed to become the compact convolutional transformer (CCT). In this investigation, CCT with Grad-CAM visualization[3] was used to find the malaria parasite.
.

Chapter 4

Implementation

4.1 Dataset

In this research, We utilized the dataset used in [22]. That dataset There are 27,558 images of cells in the dataset, and there are equal numbers of parasitized and uninfected cells. Plasmodium was present in positive samples, but other sorts of items, such as staining artifacts and contaminants, were present in negative samples. Positive samples were considered to be representative of malaria. Fig. 4.1 presents some sample images from our dataset.Here, 10% of the imagess in each dataset were utilized for testing, 20% were used for validation, and 70% were used to train the suggested model. A dataset picture is given below:

(a) Uninfected Cells

(b) Parasitized Cells

Figure 4.1: Sample Data of Malaria Cell Dataset

We have chosen to use 6000 blood cell images from this dataset for both training and testing. Because we think that employing the entire collection will cause the the models to process data too slowly and perhaps result in data loss. Additionally, some of the images that are older than 6000 images have other problems. A distribution of the dataset is given below:

Dataset	Total Images
Uninfected	13779
Parasitized	13779

Table 4.1: Data distribution

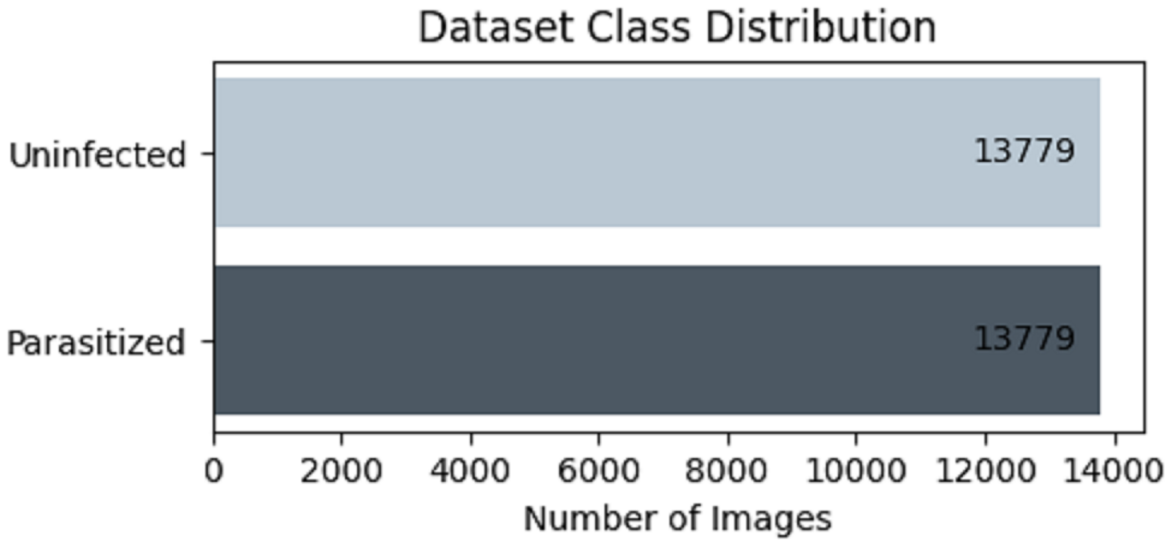


Figure 4.2: Class Distribution of Malaria Cell Dataset

It can be shown that the number of data for classes that are parasitized and uninfected is equal. As a result, there is no class imbalance in the dataset, guaranteeing that our model will never produce predictions that are favoring one class over another.

4.2 Training Set

We use the training set to teach our model architectures how the data behaves. To ensure a fair comparison, we keep the hyperparameters constant across all architectures. We teach the architectures to help them comprehend the data and its many behaviors. Our models will not be able to understand the labels or features and produce high-quality images if we do not train them. We used 70% of dataset images to train the model to be able to extract the essential elements from the data to comprehend it and forecast how the images will behave. The quality of how well they produce the photos and obtain the outcomes we require may then be compared.

4.3 Testing Set

To guarantee the validity and consistency of the data the models provided, the testing set is used to validate our findings using 10% of the dataset images. We are aware that the training and testing data varies slightly. It might result in slightly different outcomes. We will know that the architectures were correctly trained and the results acquired are valid if the findings are near or follow similar patterns. By

comparing the generated fake photos to the generated fake images from the training set, we can also utilize this to determine how similar the false images are. This will enable us to confirm the accuracy of the results we obtained and increase our faith in the model's capabilities. It will also provide us with further details about the model's performance on the provided dataset.

4.4 Validation

It is absolutely necessary to validate the model results in order to ensure that they are accurate. Our model requires a significant amount of training data in order to be trained, and the primary objective of model validation is to provide us with the opportunity to raise both the quantity and the quality of the training data. In our particular instance, twenty percent of the dataset's photos were utilized for the validation process.

4.5 Experimental setup

The simulation was carried out on a highly capable desktop that supported GPUs. This machine featured a 3.9 GHz AMD Ryzen 9 5950X processor, 64 GB of RAM, an NVIDIA GeForce RTX 3090 GPU, and a 64-bit version of Windows 10 Professional.

4.6 Performance Evaluation Procedure

During the course of our research, pictures of blood cells were obtained and then preprocessed. At the beginning, the input photos were separated into three groups: the training set, the testing set, and the validation set. After thereafter, photos received as input were processed at a resolution of 64 by 64. The input images were then subjected to both the Vision Transformer and Swin Transformer. The training simulation was executed for a total of fifty iterations for each, after which the accuracy results, as well as the precision metrics, the recall metrics, and the F1 metrics, were obtained. After that, the same metrics were recovered from the VGG-16 model after it had been trained for a total of 50 epochs. Following that, the retrieved metrics from each of the models were analyzed and compared to one another. We wanted to ensure that our comparisons were as objective as possible, so we made sure that the experimental setups and the overall architectures of the models were exactly equivalent.

$$Attention = (T_P + T_N)/(T_P + T_N + F_P + F_N) \tag{4.1}$$

$$Recall = (T_P)/(T_P + F_N) \tag{4.2}$$

$$Precision = (T_P)/(T_P + F_P) \tag{4.3}$$

Where TP = true positive indicates that a person has been accurately recognized as having malaria, TN = true negative indicates that a person has been correctly

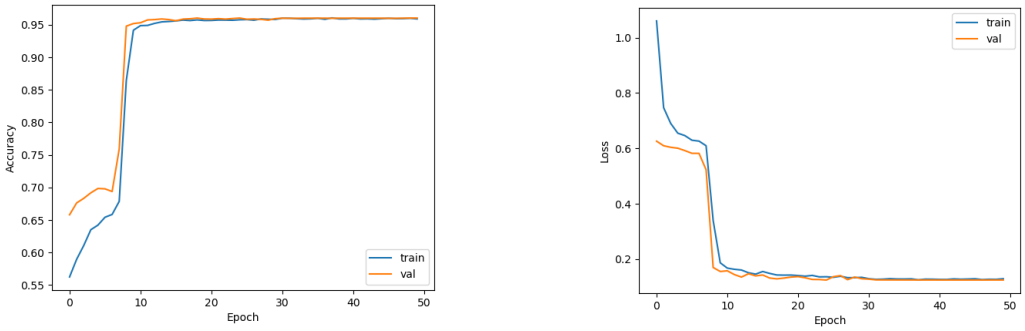
identified as not having malaria, and TP = true positive indicates that a person has been correctly identified as having malaria, The abbreviation FP stands for "false positive," which indicates that a person has been wrongly recognized as having malaria. On the other hand, the abbreviation FN stands for "false negative," which indicates that a person has been incorrectly identified as not having malaria.
In order to evaluate how well the proposed model works, numerous hyperparameters were analyzed and taken into consideration. "Adam" and "SGD" are the two optimization strategies that have been developed for deep learning so far, but "Adam" and "SGD" are the ones that have been utilized and appreciated the most. As a result, the suggested model was trained with Adam optimizers' cited' islam2022explainable' in order to demonstrate their effectiveness in identifying malaria parasites. Batch size is also crucial since it determines whether or not the model is able to learn and give outcomes that are more generic. When it comes to training, having a higher batch size can speed things up; nevertheless, having a significantly larger batch size can typically lead to poor generalization.
Calculations were done to determine the suggested model's various performance indicators, such as its precision, recall, F1-score, and accuracy. The models that use the ADAM optimizer are displayed here for a variety of different batch sizes. This illustrates that a batch size of eight was utilized in order to achieve the maximum possible results. It's possible that this is due to the fact that the model trained with Adam optimizer and larger batch sizes did not demonstrate continuous progress. In spite of the extremely high precision that the ADAM optimizer generated, the outcomes with regard to recall, F1-score, and accuracy were less than satisfactory.

Chapter 5

Result and Analysis

5.1 Result

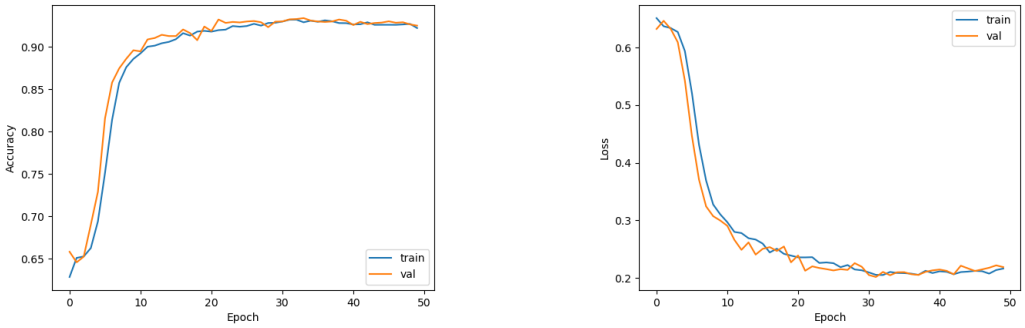
Several different hyperparameters, including optimizers, learning rates, weight decay, and batch sizes, were tested in order to undertake performance study. Confusion matrix, accuracy, precision, recall, and f1-score, among other evaluation metrics, were used to assess the performance of the models.



(a) Accuracy Curves (b) Loss Curves

Figure 5.1: Accuracy and loss graph for Vision Transformer

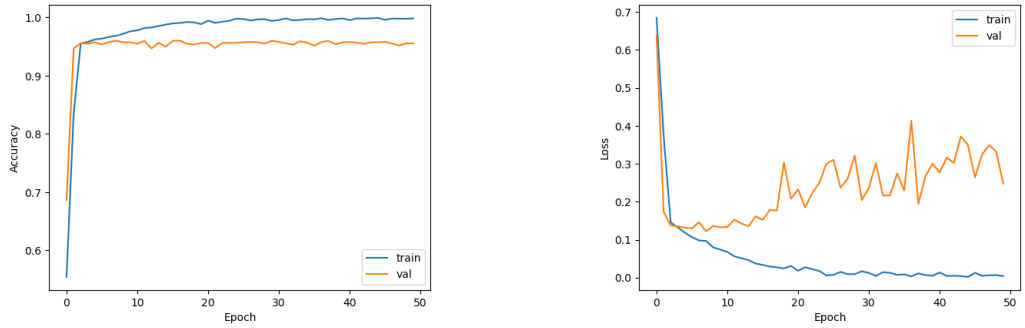
We may see a steady improvement in training performance once fifty epochs of training using Vision Transformer are complete. The validation loss approaches a plateau after roughly 20 epochs, as seen in Fig. 5.1. There was no discernible change in the outcomes after that.



(a) Accuracy Curves (b) Loss Curves

Figure 5.2: Accuracy and loss graph for Swin Transformer

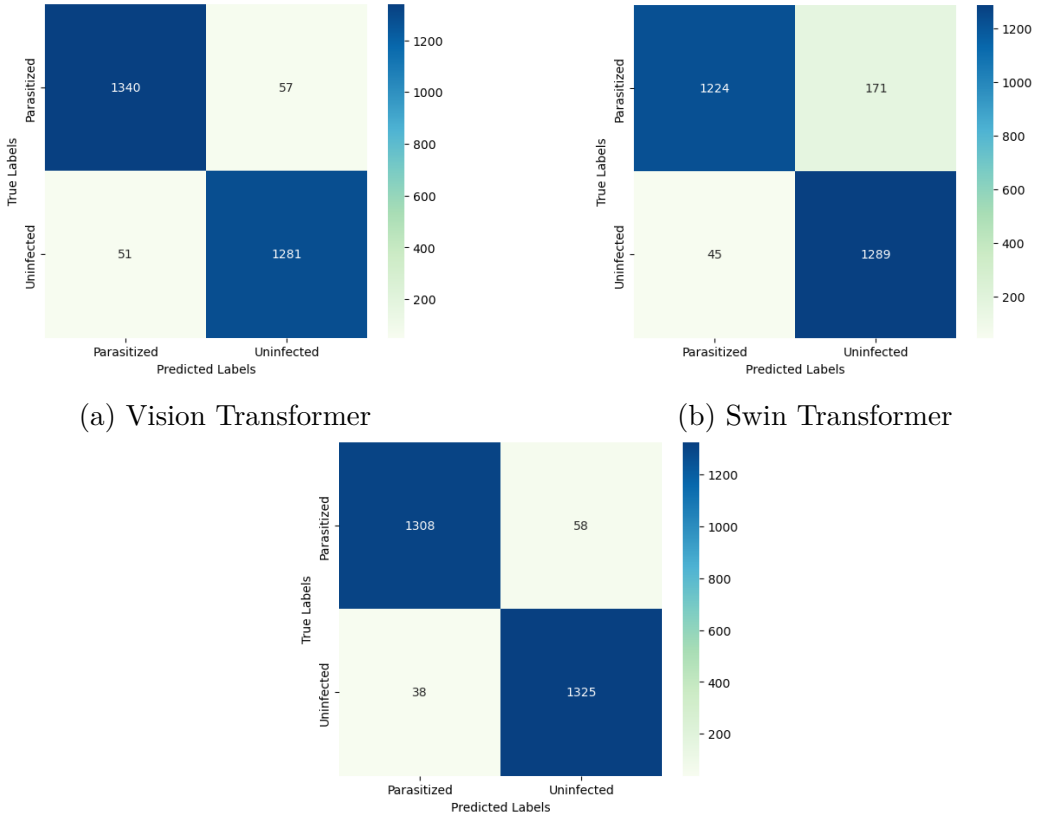
The training performance shows a definite, continuous improvement after fifty epochs of training with the Swin Transformer. As seen in Fig. 5.2, the validation loss begins to stabilize at a constant level after around 25 epochs, remains stable for the next 15 epochs, and then occasionally spikes after 42 epochs. The fact that the testing set was relatively small is one explanation for the spikes in the validation loss curve. Therefore, even small variations in the classification-misclassification scores resulted in considerable variations in the final results.



(a) Accuracy Curves (b) Loss Curves

Figure 5.3: Accuracy and loss graph for VGG-16

In the meantime, a comparison of the performance of the standard VGG-16 models and the transformer-based model shows that both models produce results with comparable minimal loss and maximum accuracy. However, as seen in Fig. 5.3, VGG models take longer than transformer-based models to achieve the maximal training plateau. The loss fluctuation is also significantly higher, which is a sign of overfitting.



(a) Vision Transformer

(b) Swin Transformer

(c) VGG-16

Figure 5.4: Confusion matrices for the three models

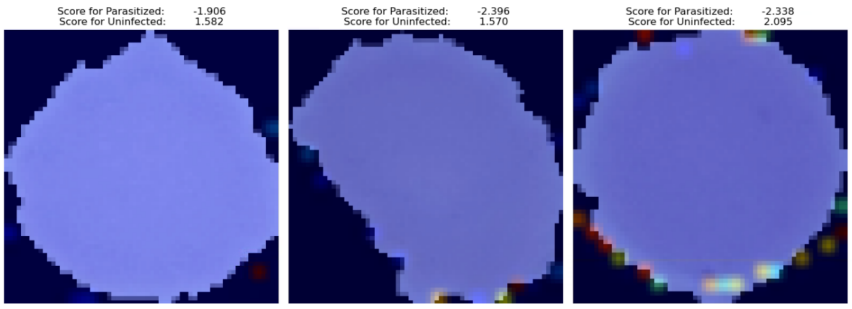
The matrices created by the Vision Transformer, Swin Transformer, and VGG-16 reflect surprisingly similar patterns, as shown in figure 5.4. Vision Transformer outperformed the other two models in terms of how well it predicted the parasitized class. However, the VGG-16 correctly identified the greatest number of examples as being in the typical class.

The two transformer-based models generally differ from one another very little in terms of training performance. The Vision Transformer, on the other hand, was longer to train than the Swin Transformer and VGG-16, taking about 22 seconds for each epoch as opposed to 15 seconds. The Vision Transformer was able to achieve excellent accuracy while keeping a more consistent training history than the other two versions.

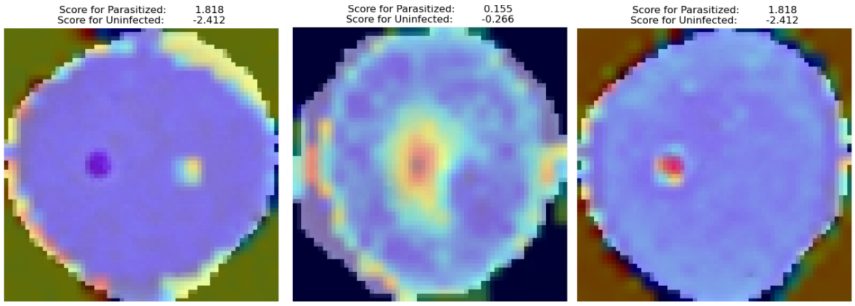
Table 5.1: Comparison of the performance on the test dataset

Models	**Accuracy**	**Precision**	**Specificity**	**Sensitivity**
Vision Transformer	0.9604	0.9592	0.9617	0.9592
Swin Transformer	0.9209	0.8774	0.9663	0.8774
VGG-16	0.9648	0.9575	0.9721	0.9575

We evaluated the implemented architectures in Table 5.1 test set scores. During the course of our test scenario, VGG16 was found to have the highest accuracy score, coming in at 96.48%, whereas Swin Transformer had the lowest accuracy, coming in at 92.09%. The accuracy rating for Vision Transformer was discovered to be 96.04%, which is considered to be about average. As a result of the scores falling into a number of discrete local maxima for each particular test, there is not much of a difference between the results themselves and the order of performance can swiftly change.Because of this, we may infer that all of the models performances are quite comparable when it comes to the results of the tests.



(a) Uninfected



(b) Parasitized

Figure 5.5: GRAD CAM Visualization

Finally, in an effort to illustrate the attention map of the Vision Transformer, Grad-Cam was used to the dataset figure 5.5 shows some of the maps that were created as a result of the correctly categorized images.The gradient weighted class activation mapping for the Vision Transformer can be very sporadically mapped throughout the images, as shown in the figure.Although there are still evidence of accurate activation mapping where the parasites are marked, the area also has sporadic activation mapping. This sparse distribution of attention maps is typically brought about by the fact that the mapping merely conveys information about the places the network paid attention to, ignoring the areas that were really employed to determine the final categorization. Consequently, the mapping around the images is sparse.

Chapter 6

Conclusion and Future Work

The goal of this thesis was to employ a heavily hyped computer vision and image recognition model, the vision transformer model, to identify malaria parasites in human blood cells enabling rapid malaria disease diagnosis. We have highlighted the advantages of vision transformers in image categorization versus formerly utilized convolutional neural networks including VGG16. To find the best transformer model, we then used a specific variant of vision transformer, the swin transformer and compared their performance.The accuracy of the vision transformer variant, swin transformer is lower than that of the vision transformer, although only by a tiny margin (95% to 92%). Overall, it appears that the transformer model outperforms the VGG16 model. As previously indicated, the vision transformer diagnosed the patient more precisely than that of the VGG16.The vision transformer model proposed in this thesis can detect malaria parasites at the same or slightly greater accuracy than previously mentioned methods while using fewer resources. Nevertheless, it is the greatest benefit over standard convolution neural networks, such as the previously discussed VGG16, is speed. While comparing to the VGG16, the vision transformer gets around a quarter of the speed advantage in training phase.While the model has gained momentum in terms of computational speed, it may be enhanced further by utilizing newer, high-processing environments that are better equipped for applications involving deep learning. Second, when a much larger input is fed into the network, vision transformers achieve greater detection accuracy. We predict that the classification of the vision transformer model would dramatically improve if there was a much larger dataset, perhaps numbering in the millions, as opposed to the mere thousand used in this thesis.Vision transformers are a relatively recent trend in computer-aided image recognition. As a consequence of this, it has garnered a negligible amount of attention in the field of medical diagnostics, which has traditionally relied on alternative methods such as CNNs. We exhibited the comparison and the benefits it offers in categorizing cell pictures, which is the most important component in correctly diagnosing malaria. This is because it has a tendency to perform better. The learning was also validated with the use of Grad-CAM visualization, which was utilized to examine the trained model. In the study that was suggested, employing the transformer model allowed for accuracy, precision, and recall to be achieved, respectively. Comparisons were made between the recommended model and the numerous SOTA initiatives for parasite detection in malaria. The results of the Vision Tranformer's identification of malaria parasites were superior to those of any preceding investigation. The primary focuses of our

upcoming research will be on working some other Transformer based approach to upgrade the accuracy of the models.

Bibliography

[1] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[2] Q. Guan, Y. Wang, B. Ping, *et al.*, "Deep convolutional neural network vgg-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: A pilot study," *Journal of Cancer*, vol. 10, no. 20, p. 4876, 2019.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[4] A. Marefat, J. H. Joloudari, and M. Rastgarpour, "A transformer-based algorithm for automatically diagnosing malaria parasite in thin blood smear images using mobilevit," EasyChair, Tech. Rep., 2022.

[5] Y. G. Gezahegn, Y. H. G. Medhin, E. A. Etsub, and G. N. G. Tekele, "Malaria detection and classification using machine learning algorithms," in *International Conference on Information and Communication Technology for Develoment for Africa*, Springer, 2017, pp. 24–33.

[6] A. Alqudah, A. M. Alqudah, and S. Qazan, "Lightweight deep learning for malaria parasite detection using cell-image of blood smear images.," *Rev. d'Intelligence Artif.*, vol. 34, no. 5, pp. 571–576, 2020.

[7] P. Pandit and A. Anand, "Artificial neural networks for detection of malaria in rbcs," *arXiv preprint arXiv:1608.06627*, 2016.

[8] M. R. Islam, M. Nahiduzzaman, M. O. F. Goni, *et al.*, "Explainable transformer-based deep learning model for the detection of malaria parasites from blood cell images," *Sensors*, vol. 22, no. 12, p. 4358, 2022.

[9] K. Fuhad, J. F. Tuba, M. Sarker, *et al.*, "Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application," *Diagnostics*, vol. 10, no. 5, p. 329, 2020.

[10] A. Khan, K. D. Gupta, D. Venugopal, and N. Kumar, "Cidmp: Completely interpretable detection of malaria parasite in red blood cells using lower-dimensional feature space," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.

[11] 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.

[12] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," *Translational Research*, vol. 194, pp. 36–55, 2018.

[13] F. J. P. Montalbo and A. S. Alon, "Empirical analysis of a fine-tuned deep convolutional model in classifying and detecting malaria parasites from blood smears," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 15, no. 1, pp. 147–165, 2021.

[14] M. K. Sagor, S. M. Dipto, I. Jahan, S. Chowdhury, M. T. Reza, and M. A. Alam, "An efficient deep learning approach for detecting lung disease from chest x-ray images using transfer learning and ensemble modeling," in *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, IEEE, 2021, pp. 1–5.

[15] K. Sriporn, C.-F. Tsai, C.-E. Tsai, and P. Wang, "Analyzing malaria disease using effective deep learning approach," *Diagnostics*, vol. 10, no. 10, p. 744, 2020.

[16] V. Magotra and M. K. Rohil, "Malaria diagnosis using a lightweight deep convolutional neural network," *International Journal of Telemedicine and Applications*, vol. 2022, 2022.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] S. Rajaraman, S. Jaeger, and S. K. Antani, "Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images," *PeerJ*, vol. 7, e6977, 2019.

[19] Z. Liu, H. Hu, Y. Lin, *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 009–12 019.

[20] D. Bibin, M. S. Nair, and P. Punitha, "Malaria parasite detection from peripheral blood smear images using deep belief networks," *IEEE Access*, vol. 5, pp. 9099–9108, 2017.

[21] T. Jameela, K. Athotha, N. Singh, V. K. Gunjan, and S. Kahali, "Deep learning and transfer learning for malaria detection," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[22] S. Rajaraman, S. K. Antani, M. Poostchi, *et al.*, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, e4568, 2018.

[23] Z. May, S. S. A. M. Aziz, *et al.*, "Automated quantification and classification of malaria parasites in thin blood smears," in *2013 IEEE International Conference on Signal and Image Processing Applications*, IEEE, 2013, pp. 369–373.

[24] N. Jain, A. Chauhan, P. Tripathi, S. B. Moosa, P. Aggarwal, and B. Oznacar, "Cell image analysis for malaria detection using deep convolutional network," *Intelligent Decision Technologies*, vol. 14, no. 1, pp. 55–65, 2020.

[25] S. Khademi, S. Heidarian, P. Afshar, *et al.*, "Spatio-temporal hybrid fusion of cae and swin transformers for lung cancer malignancy prediction," *arXiv preprint arXiv:2210.15297*, 2022.

[26] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*, Springer, 2022, pp. 272–284.

[27] Y. Dong, Z. Jiang, H. Shen, *et al.*, "Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells," in *2017 IEEE EMBS international conference on biomedical & health informatics (BHI)*, IEEE, 2017, pp. 101–104.