Vision Transformer (ViT) Approach in Computer Aided Diagnosis of Acute Lymphoblastic Leukemia

by

Sifatul Amin
18101144
MD. Samin Jawed
18101085
MD. Rejuan Rashed Raj
18301165
MD. Sabbir Ahmed Saimoon
18101083
MD. Rakibuzzaman Rayhan
18101082

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
Brac University
January 2023

Declaration

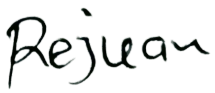
It is hereby declared that

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:

Sifatul Amin
18101144

MD. Samin Jawed
18101085



MD. Rejuan Rashed Raj
18301165

MD. Sabbir Ahmed Saimoon
18101083

MD. Rakibuzzaman Rayhan
18101082

Approval

The thesis titled "Vision Transformer(ViT) approach in computer aided diagnosis of Acute Lymphoblastic Leukemia" submitted by" submitted by

1. Sifatul Amin(18101144)
2. MD. Samin Jawed(18101085)
3. MD. Rejuan Rashed Raj(18301165)
4. MD. Sabbir Ahmed Saimoon(18101083)
5. MD. Rakibuzzaman Rayhan(18101082)

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

Examining Committee:

Supervisor:
(Member)

Annajiat Alim Rasel
Digitally signed by Annajiat Alim Rasel
DN: cn=Annajiat Alim Rasel, o=Brac University, ou=CSE Department, email=annajiat@bracu.ac.bd, c=BD
Date: 2023.01.14 23:04:00 +06'00'

Annajiat Alim Rasel
Senior lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Tanzim Reza
Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Chair)

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

Head of Department:
(Chair)

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

Abstract

Blood cancer is a serious and potentially deadly type of cancer that affects the production of blood cells in the body.Leukemia, lymphoma, and myeloma are the three primary kinds of blood cancer.Leukemia, which is the most common and deadly type of blood cancer, is characterized by the abnormal and unexpected development of white blood cells (leukocytes) in the bone marrow. Leukemia comes in two primary varieties: acute and chronic. Acute leukemia progresses more quickly and is more common in children, while chronic leukemia progresses more slowly. Early detection of leukemia is important for proper treatment, as it can be fatal if not treated promptly. One method of detecting leukemia is through imaging, which is quick and inexpensive and does not require specialized equipment or laboratory tests. However, manual classification of leukemia cells by hematologists can be time-consuming and prone to errors. In recent years, the preferred technique for vision application is convolutional neural networks (CNNs).CNN have demonstrated their effectiveness in automatically classifying medical images. However, their limited local receptive field can prevent them from learning global context information. An alternative to CNNs that has shown promise is the Vision Transformer (ViT), which uses self-attention between image patches to process visual information. However , Vit does not work very well without a large dataset so we are using the ISBI 2019 data set, a dataset of 10000+ images and this data set needs more polishing, we're not just suggesting a transformer architecture for diagnosing ALL; we're also laying the groundwork for its polishing and sharing every piece of code we've used in our research. Our Vit model produces an accuracy of 81.5%, and shows how it has potential to reach new heights. The suggested approach has the ability to accurately differentiate between cancer cells knows as B-lymphoblast cells and normal cell known as B-lymphoid precursors and can be utilized as an efficient technique for assisting in the effective discovery of acute lymphoblastic leukemia through computer assistance.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our Supervisor Mr. Annajiat sir for his kind support and advice in our work, and Mr. Tanzim sir as they helped us whenever we needed help.
And finally to our parents without their support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

Hematologic malignancy, commonly referred to as blood cancer, is a specific type of cancer that impacts the bone marrow and blood cells. It is characterized by the unchecked proliferation and division of aberrant blood cells, which may affect the formation of healthy blood cells and harm the immune system, red blood cells, and bone marrow function [21]. Leukemia, Lymphoma, and Myeloma are the three main kinds of blood cancer, which all begin in the bone marrow [8]. These cancerous cells, also known as aberrant blood cells, can interfere with the normal functions of blood, such as preventing excessive bleeding. There are several types of blood cancer such as:

1. Leukemia: A bone marrow malignancy that affects the blood-forming cells. Uncontrolled proliferation and division of aberrant white blood cells, which could accumulate in bone marrow and prevent the production of healthy blood cells, are the hallmarks of this condition. Acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), and chronic myeloid leukemia are the additional four primary subtypes of leukemia (CML).[22].
2. Lymphoma: A lymphatic system malignancy that affects the immune system. Hodgkin's lymphoma and non-lymphoma Hodgkin's are the two main kinds of lymphoma (NHL). NHL is characterized by the presence of a different type of cell than Hodgkin's lymphoma where Reed-Sternberg cells are present, as a characteristic.
3. Myeloma: A cancer of plasma cells, which are a type of white blood cell. Myeloma is characterized by the overproduction of abnormal plasma cells, which can accumulate in the bone marrow and form tumors in bones throughout the body.
4. Myelodysplastic Syndromes (MDS): A group of disorders which causes an inadequate production of platelets, white blood cells, and red blood cells by the bone marrow. It is a precancerous condition that can progress to acute myeloid leukemia (AML) over time.
5. Myeloproliferative Neoplasms: Is a group of blood cancers characterized by the overproduction of cells in the bone marrow, which results in an unnatural rise in the quantity of red blood cells, white blood cells, or platelets in the bloodstream. Each of these types of blood cancer has its own set of characteristics, causes, and treatment options. Figure 1 shows different types of leukemia. The type and stage of blood cancer will influence the treatment approach. In this paper, we will just

work with only a subtype of leukemia and that is Acute lymphoblastic leukemia (ALL). Acute lymphoblastic leukemia (ALL) is a serious type of blood cancer that affects both children and adults. According to the estimates of the World Health Organization (WHO), around 232,000 new cases of leukemia were diagnosed and 191,000 deaths occurred due to leukemia in 2018 worldwide. The deaths caused by ALL specifically would be a part of that number, but the exact statistics are not provided by the WHO. It's important to note that over the years, the survival rate for ALL has increased significantly thanks to advances in treatment and research. With early diagnosis and appropriate treatment, many people with ALL can go into remission and lead healthy lives. The third largest cause of death is blood cancer. Blood cancers have a 70% overall survival rate; however, the survival rates of various subtypes vary greatly. Research that was released at the end of the previous year examined the results of 3,377 blood cancer patients who had contracted the coronavirus worldwide. It was shown that 34% of individuals who did not have blood cancer but were hospitalized with the coronavirus died, as opposed to 22% of those who were. This means that among hospital patients, persons with blood cancer appear to have an approximately 50% higher risk of dying from a coronavirus than people without blood cancer.
Let's talk about CNN, In image processing and computer vision tasks, convolutional neural networks (CNNs), a form of deep learning method, are frequently utilized. They have also been used to help with the diagnosis of blood cancers like leukemia. CNN's can be trained to analyze images of blood cells, such as those obtained through a microscope, and identify patterns that are associated with different types of leukemia. This can help to automate the process of diagnosing the disease and reduce the dependence on subjective human interpretation. For example, researchers have used CNNs to analyze microscopic images of blood cells from patients with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), and the results were able to differentiate the two types of leukemia with high accuracy. Also, CNN's can be used to classify the images based on different subtypes of leukemia, like B-cell ALL and T-cell ALL, or different genetic mutations which can be associated with specific outcomes or treatment responses. In addition, with the advancements in technology, CNN can indeed be used to analyze a variety of data sources, including images from microscopes, flow cytometry, Cytogenetics, Next-Generation-Sequencing, and more. With the integration of this data, the model can have higher diagnostic accuracy.

However, a more recent deep learning architecture called Vision Transformer (ViT) has been shown to outperform CNNs in some image classification tasks, including the diagnosis of blood cancers. CNN's use a series of convolutional layers to extract features from images, while ViT uses a self-attention mechanism to analyze the entire image at once. This allows ViT to better capture global dependencies in images and make more accurate predictions. Additionally, the Vision Transformer architecture can handle a large number of parameters and can be trained on a large dataset which can increase the accuracy of the model. Also, by using ViT, the feature extraction process is not limited to just the spatial relationship of the pixels but also the contextual relationship between the pixels. The figure1.1 below show types of leukemia cancer taken from [7].

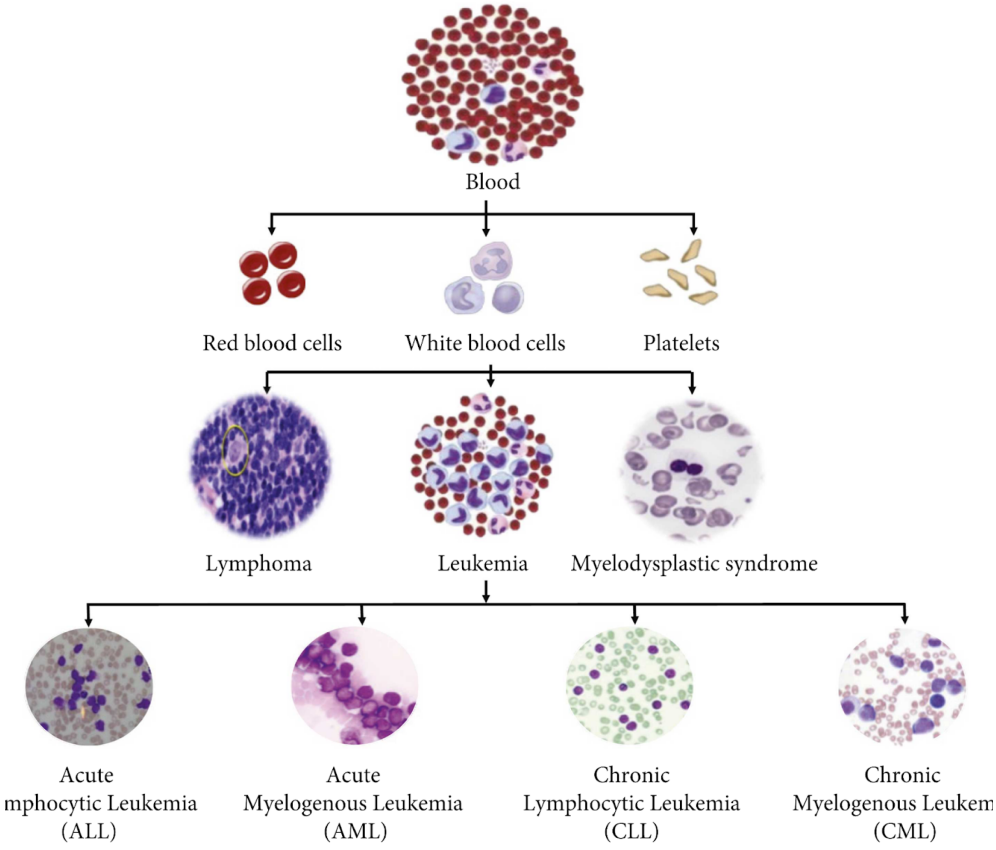


Figure 1.1: Types of Leukemia Blood Cancer

1.2 Research Objectives

Our research objective is to use Vision Transformer (ViT) in Acute Lymphoblastic Leukemia (ALL) diagnosis. The main objective of using Vision Transformer (ViT) models is to improve the accuracy and objectivity of the diagnostic process, compared to traditional methods or even convolutional neural networks (CNNs). ViT models can process images in a different way than CNNs and have the ability to better capture global dependencies in images and make more accurate predictions. They also have the capability to handle a large number of parameters and can be trained on a large dataset which can increase the accuracy of the model. Additionally, by using ViT the feature extraction process is not limited to just the spatial relationship of the pixels but also the contextual relationship between the pixels, which can be helpful for the diagnosis of ALL. Moreover, with technological advancements, ViT can be used on a wide range of data sources such as microscopy images, flow cytometry, Cytogenetics, Next-Generation-Sequencing, and more. With the integration of this data, the model can have higher diagnostic accuracy compared to traditional methods. Alongside, medical sectors need more research on ViT if this methodology is actually a good option for medical diagnosis and thus contributes to the medical sector worldwide.

1.3 Problem Statement

There are several challenges currently faced in the diagnosis and treatment of Acute Lymphoblastic Leukemia (ALL):
1. Late diagnosis: ALL often has few or non-specific symptoms in its early stages, which can make it difficult to diagnose in a timely manner. This can lead to a delay in treatment and a poorer outcome for the patient.
2. Risk stratification: ALL is a diverse disease, with varying outcomes based on the genetic and molecular characteristics of the tumor. However, accurately identifying high-risk patients can be challenging, and there are no universally accepted criteria for defining risk groups.
3. Relapse: Despite initial remission with standard treatment, some patients will experience a relapse. The development of resistance to chemotherapy or targeted therapy is a major challenge in the treatment of relapsed or refractory ALL.
4. Toxicity: Treatment for ALL often involves high doses of chemotherapy and other drugs, which can cause significant side effects and long-term health problems. This can be especially challenging for older adults and children, who may be more vulnerable to the toxic effects of treatment.
5. Precision medicine: Advances in genetics and molecular biology have led to the identification of specific genetic mutations that drive the development of ALL. However, targeting these mutations with targeted therapies can be challenging. Additionally, it is still a question of how these genetic mutation findings can be translated into personalized medicine.
6. Reliance on bone marrow transplantation: Bone marrow transplantation (BMT) is a powerful treatment option for certain subtypes of ALL, but this is linked to higher fatality as well as instability and is not always feasible. Therefore, there is a need to develop less invasive and less toxic alternatives to BMT.
7. Cost: The cost of treatment for ALL can be high, and access to care can be limited for some patients, especially those without adequate insurance coverage.

Researchers are actively working on finding new diagnostic methods, therapies, and novel combination therapies to overcome these challenges and improve the outcomes for patients with ALL. The need for improved diagnostic accuracy and objectivity to improve patient outcomes and reduce the dependence on subjective human interpretation has to be taken into consideration. Convolutional neural networks (CNNs) have demonstrated tremendous capability in the automatic categorization of US images and have become the method of choice in vision applications over the past 10 years. Successful, one of the main limitations of CNNs is that their ability to learn global context is restricted by the size of the local receptive field. This is because the local receptive field only looks at a small, local portion of the input data at a time, and the information outside of this local region is not taken into account when computing the output feature maps. As a result, CNN's may struggle to capture global patterns and dependencies in the data, which can be important for certain tasks. This is in contrast to models such as vision transformers (ViTs), which use self-attention layers instead of convolutional layers. These layers allow the model to attend to different parts of the input data at different positions in the processing, which can help the model capture global context information more effectively.

Chapter 2

Literature Review

Using computer vision technology to aid in identifying illnesses is an area of active research in recent years. One key method in this technology is using deep learning for image recognition. Among the neural networks commonly used in deep learning, convolutional neural networks (CNN) stand out for their strong abilities to learn and adapt on their own, as well as generalize to new data. In contrast to traditional image recognition techniques that rely on manually extracting and classifying features, CNNs need the image data alone as input, allowing the model to classify the images through its capabilities of learning on its own [20][4].
Nahid et al. employed a multi-channel convolutional neural network to detect and diagnose pneumonia through chest radiographs. The resulting classification accuracy rate was 97.92%, which was found to be a highly dependable detection method as per the reference [15].
Daoud et al. utilized cardiac ultrasound images, a combination of computational intelligence methods for obtaining imaging attributes, and manual extraction methods used to categorize breast tumors. 96.1% of classifications were accurate on average, that indicated cancer in the breast can be correctly detected using an ultrasound picture of a breast, as per reference [10].

Yang et al. in their paper, bladder cancer was identified using deep learning and the classification accuracy rate in real-world use was 83.36%. This demonstrated the effectiveness of deep learning in diagnosing bladder cancer because its accuracy was on par with that of medical professionals., as per the reference [19].

Similarly, some researchers are utilizing computer vision to detect leukemia. According to the reference, Ahmed et al. used convolutional neural networks and machine learning techniques to identify four forms of leukemia, with the greatest accuracy being 88.25% [2].

Boldú et al. in there, a machine learning approach was presented for identifying acute leukemia centered on peripheral blood pictures. They used color morphology in mathematics and clustering for segmenting the pictures, and later applied methods for machine learning to categorize six different cell kinds. The diagnostic accuracy for leukemia was 94% while the detection accuracy for cell classification was 85% + as per the reference [3].

Kasani et al. successfully distinguished leukemic B-lymphoblastic cells from normal B-lymphoid progenitor cells. using a mix of the two models NASNetLArge and VGG19. They also demonstrated, in the reference, that the ensemble of classifiers performed significantly better than a single network [13].

In recent years, transfer learning has gained popularity as a method for classifying medical images. Alshazly et al. achieved an accuracy rate of 92.9% on the COVID-19-CT dataset by using this approach to train chest CT scans for identifying COVID-19 patients. They also used graphical representations, according to the source [16], to properly articulate the prediction performance.

Brodzicki et al. utilized transfer learning techniques to train a convolutional neural network that was able to classify the cytotoxicity of Clostridium difficile bacteria with a high accuracy of 93.5% on 369 images and demonstrated exceptional performance [9].
Nahzat et al. introduced a method for classifying white blood cells using a convolutional neural network. They experimented with various optimizers and found that RMSprop produced the highest performance. Their suggested model included a flattened layer, fully connected layers, a final fully connected layer using the SoftMax activation function, and five convolutional blocks for extracting features. The first block used Conv2D and the subsequent blocks used SeparableConv2D. They obtained a high F1 score of 99%, strong recall, and an accuracy of 99.5% [25].

These studies all utilize convolutional neural networks (CNNs) as their model. However, an alternative model known as a vision transformer, which depends on the transformer's architecture, utilizes a self-attention mechanism that is different from CNN's. The transformer structure was originally developed for natural language processing (NLP) but was later adapted to computer vision. Compared to CNNs, the transformer-based model demonstrated improved performance in image classification.[17][18].

Here are some research reviews on several implementations of ViT structure in different sectors:

Zhou et al. discuss a novel method for captioning images using a deep encoder-decoder model built on a sparse Transformer framework. This method uses multiple levels of image features and self-attention to focus on both minute details and the broad context in the encoder and then uses the most crucial pieces of this information in the decoder to create the caption. A new model called Local Adaptive Threshold is introduced, which is able to efficiently parse the attention matrix and provide more focused attention than the standard Transformer model. The article provides details about the changes made to the self-attention module and the design of the image captioning model, and it is evaluated using MSCOCO and Flickr30k datasets [14].

The authors propose a new method for image compression that combines patch-based processing with vision transformers and a context model called TransContext. They suggest using transformers in the foundational layer of the network and incor-

porating residual coding to achieve compression at different bit rates. 40,000 photos from the COCO-2014 collection served as the test set for their method, while the Kodak PhotoCD dataset 2 and the BSD 100 test datasets were used to assess the performance [24].

The authors of the paper present AnoViT, an encoder-decoder model that uses a vision transformer to locate and identify image abnormalities. The model captures normal information by also learning the relationship between image patches globally. They propose two methods: VIT-BASED ENCODER-DECODER, which uses an encoder with a Vision Transformer (ViT) and a decoder with a convolutional layer, and ANOMALY DETECTION AND LOCALIZATION, which uses reconstruction error to detect and locate anomalies. The proposed model was tested using the MNIST, CIFAR10, and MVTecAD datasets to evaluate the performance of anomaly detection and localization [23].

Vishwani et al. in their publication demonstrate the effectiveness of self-attention as the sole necessary component and their experiments on two machine translation tasks reveal that these models have better quality, can be processed in parallel more easily, and need less time for training [1].

Chapter 3

Materials and Methods

3.1 DATASET

The ISBI 2019 dataset [12], comprising 10,661 microscopic white blood cell images. Blood samples were taken from 73 individuals to construct the diagnostic model. This dataset includes 7,272 images of B-lymphoid leukemia cancer cells from 47 cancer patients and 3389 healthy cell images from 26 different healthy individual. These cells were separated from the microscopic photographs that means the white blood cells (WBC) were pre segmented from other cells, Needs mentioning that each cell image is genuine. This dataset is an imbalanced set. Using 'Difference Enhancement-Random Sampling' (DERS) would solve that problem and thus give better results but we are not implementing that in this paper. The Dataset is divided into three folds each containing all-leukemia images and hem-healthy images. We have merged all the folds for training purposes.Some of the image examples from the dataset is shown below on figure 3.1.

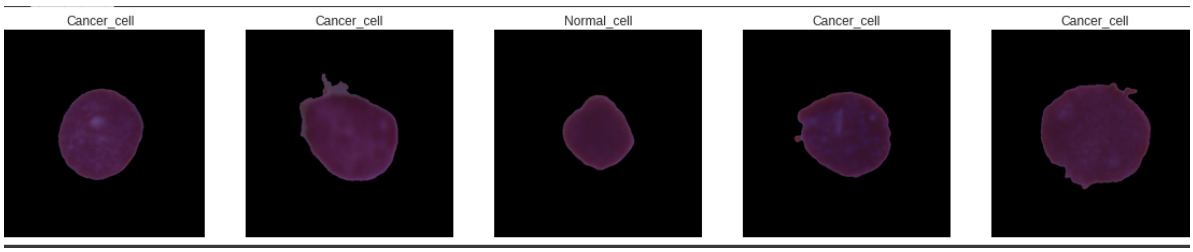


Figure 3.1: Microscopic Image examples from the Database

A qualified oncologist will comment on the label of the picture since, as seen in Figure 2, the two cells' shapes resemble one another somewhat. Labels for photos of healthy cells are positive samples, whereas labels for images of malignant cells are negative samples.

3.2 Data Pre-Processing

It is crucial to scale the dataset in computer vision and image processing. The size of the items in an image may vary widely, and the characteristics that are used

to describe the picture can also have variable scales, especially in tasks like object recognition and image classification. Scaling the features of an image can enhance the performance of the image processing algorithms by decreasing the impact of variations in scale. For example, normalizing an image's pixel values to fall between the range of 0 and 1 can guarantee that changes in illumination won't have an impact on the algorithm. In addition to enhancing algorithm performance, dataset scaling can make the algorithms more flexible to changes in the input data. For example, if an image processing algorithm is trained on an image dataset that is not scaled, it might not perform well on new images that have different lighting conditions or different scales of objects. The selection of the scaling method to use is dependent on the specific dataset and the machine learning algorithm being used. We used normalization in our case.

3.2.1 Image scaling & splitting:

Our Dataset consists of 10661 segmented white blood cell microscopic images. All the images are in 450 x 450 pixel resolution. To reduce the input load of the model we have down scaled all the images into 200x200 pixels which will reduce training time complexity.

The most important and unique part of the vision transformer is dividing images into patches. In our research, we have implemented spliting techniques to all the images and produced tokens. Every image was split into 64 patches. Each patch size is set as 25x25 pixels, shown in figure 3.2.

The tokenization process is crucial for Vision Transformer models, as it allows the model to treat images in a similar manner to how it processes text. This enables the model to utilize the transformer architecture for various tasks like image classification and object detection.

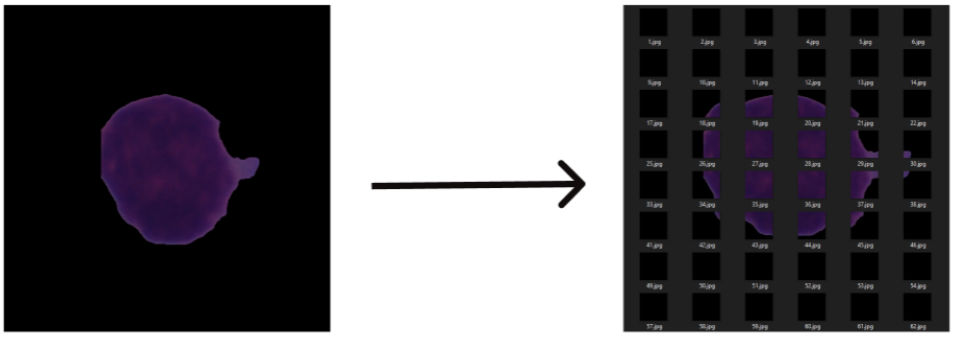


Figure 3.2: Split an image into 64 patches

3.2.2 Patch & Position Embedding:

The standard Transformer accepts input as a 1D series of token embedding. Patch & Position Embedding. Pictures were molded into a series of 2D patches that were flattened in order to handle 2D images.
To keep track of positions, position embedding is added to patch embedding. These embedding can either indicate a feature's position in a 1-dimensional flattened sequence or a feature's position in 2-dimensional space in computer vision.
Figure 3.3 shows a basic demonstration of our workflow.

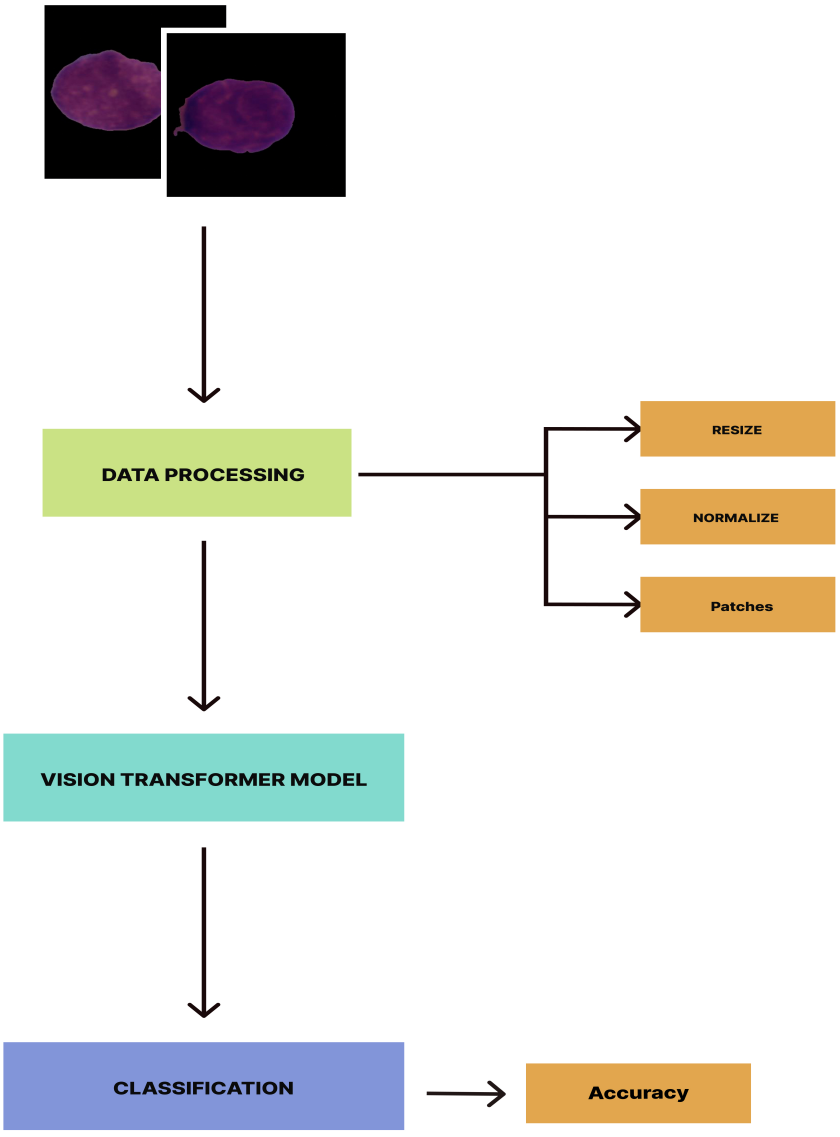


Figure 3.3: Work Flow Diagram

3.3 Classification Model

This section presents some of the models that inspired us to reach this research point.

3.3.1 CNN

A convolutional neural network (CNN) is built using various layers such as convolutional layers, an activation function, pooling layers, and fully connected layers. Convolutional, activation, and pooling layers make up the feature extraction layer in CNN classification models, which is utilized to extract features. In figure 3.4 convolution function is shown. The pooling layer is a downsampling technique that lowers the retrieved features' dimensionality while retaining crucial data. As seen in the following equation, the convolutional layer is the foundation of a CNN [5] [6].

$$y(t) = \int_{-\infty}^{\infty} x(p)\, h(t-p)\, \mathrm{d}p = x(t) * h(t).$$

Figure 3.4: Convolution function of CNN

3.3.2 ResNeXt

ResNeXt (Residual Network with Extreme Inverted Bottlenecks) is a convolutional neural network architecture that was introduced in 2016 by researchers at Facebook AI Research. It is an extension of the popular ResNet architecture and is designed to improve the performance of image classification tasks by increasing the model's capacity while reducing computational complexity. The main innovation in ResNeXt is the use of "inverted bottlenecks" in the network architecture. An inverted bottleneck is a combination of a 1x1 convolutional layer, which reduces the number of channels, followed by a 3x3 convolutional layer, which increases the number of channels. The use of this structure allows ResNeXt to increase the model's capacity while reducing computational complexity. ResNeXt also uses a technique called "cardinality", which is a measure of the number of parallel paths in a network. In ResNeXt, the cardinality is increased by splitting the feature maps into different groups and applying different convolutional filters to each group. This allows the model to learn more diverse features and improves its ability to generalize to new data.

3.3.3 Vision Transformer

The Vision Transformer, or ViT, is a classification approach that applies a Transformer-like pattern to certain portions of the image. By partitioning an image into rectified size patches, linearly embedding all these, and then integrating position embedding, a sequence of vectors is produced. then finally supplying the completed vector sequence to a conventional Transformer encoder. A further trainable "classification

token" is inserted into the sequence as part of the standard classification procedure. The Transformer architecture has gained widespread acceptance in the field of natural language processing, but its use in computer vision is somewhat limited. In computer vision, convolutional networks are either paired with attention mechanisms or parts of their components are substituted while preserving the overall structure of the network. However, this study shows that it is not necessary to rely on CNNs and that pure Transformer models can perform well on image classification tasks when applied directly to sequences of image patches. When tested on various small to medium-sized image recognition benchmarks such as ImageNet, CIFAR-100, VTAB, etc., the Vision Transformer (ViT) model achieves excellent results with significantly less computational resources required for training [11].

Here is a brief overview of how a ViT works:
1. The transformer design divides the input picture into a grid of smaller patches, which are each processed as a separate "token". These tokens are processed by a series of self-attention layers, which allow the model to attend to different parts of the image at different positions in the processing.
2. The output of the self-attention layers is passed through a series of fully-connected layers, which learn to classify the input image based on the features extracted by the self-attention layers.
3. The final output of the ViT is a set of predictions or class scores for the input image.

As seen in the formula, Vit has numerous threads of attention, which are a kind of self-attention framework that allows the system to focus on different informative components. Using the multi-head attentive (1)-formula (3).
In our binary cancer cell detection we have used 'sigmoid' activation in dense layers. To calculate model loss we have used 'binary crossenropy'.

Major layers of ViT:

1. LN: Layer Normalization let the input data in the given batch to be autonomously normalized. Batches of just about any size can use layer normalization because it is independent of batch size.
2. MHA: A Multihead Self-Attention Layer (MHA) is a component of the Vision Transformer architecture which allows the model to attend to multiple regions of an image simultaneously. It applies multiple attention mechanisms in parallel, each with its own set of parameters. Formula of MHA is shown in Figure3.5. The MHA layer is based on the self-attention mechanism, which allows the model to weigh the importance of different regions of an image when making a prediction.

3. MLP: Multi-Layer Perceptron (MLP) is a type of artificial neural network which is made up of several layers of artificial perceptrons or neurons. It is a feedforward network, which means that information only moves from the input to the output layer in a straight line. The Gaussian Error Linear Unit (GELU) is an activation function of a neuron in neural networks. It is defined as: f(x) = x * (x), where is the cumulative density function of a standard normal distribution. It is used to weigh the input layer.

$$
\begin{aligned}
Q_i &= QW_i^Q, \\
K_i &= KW_i^K, \\
V_i &= VW_i^V, i = 1, \ldots, 8, \\
\text{head}_i &= \text{Attention}\left(Q_i, K_i, V_i\right), \quad i = 1, \ldots, 8, \\
\text{MultiHead}\left(Q, K, V\right) &= \text{Concact}\left(\text{head}_1, \ldots, \text{head}_8\right) W^o.
\end{aligned}
$$

Figure 3.5: Multi Head Self-Attention Layer

The terms Q, K, V, and K in these calculations stand for the query vector, key vector, value vector, and weight matrix, respectively.

The vision transformer approach typically includes the layer of linear embedding. The image is divided into several patches, each of which is reconfigured into a (1D) one-dimensional tensor. Location embedding and class embedding are combined after the patch embedding process is complete and fed into the transformer encoder. Following that, an MLP unit structural system composed of an activation function and a fully connected layer is used to route the transformer encoder's output. The GELU (Gaussian Error Linear Unit) activation function shown in Figure 3.6.

$$
\text{GELU}\left(x\right) = 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right).
$$

Figure 3.6: GELU activation function

The illustration of a basic vision transformer model is depicted in Figure 3.7.

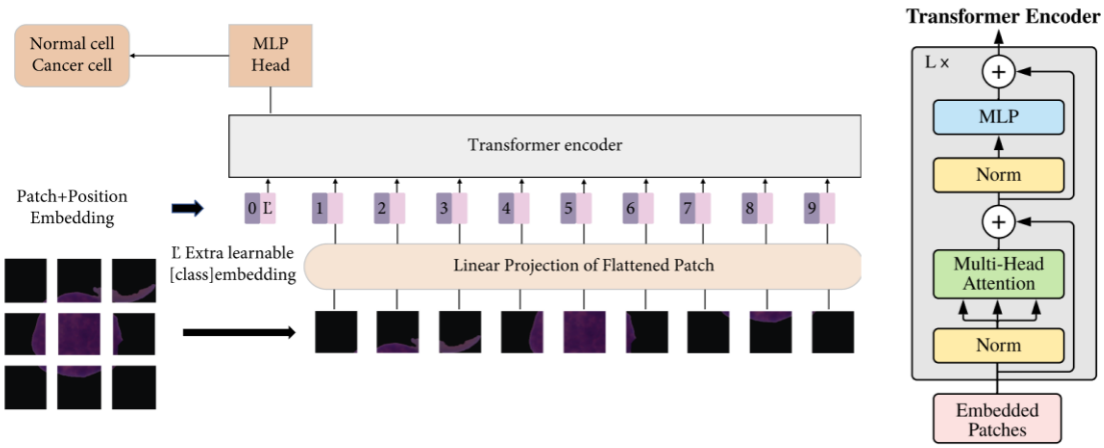


Figure 3.7: Transformer Model & Encoder

Chapter 4

Result Discussion

4.1 Experiment Environment

In our paper, we have implemented a Base model of ViT which consists of 12 layers with 86M parameters.
The hardware environment employed in this paper's experimental setup comprises Google Colab cloud GPU. Python 3.8 environment was used. We have used tensorflows 2.11 as our machine learning library. The microscopic blood image dataset takes around 10 GB and our training time was 21 hours. The ViT-Base model information is given below in Figure 4.1.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M

Figure 4.1: ViT-Base Model

4.2 Performance Metrics

Several metrics may be used to assess an image classification model's effectiveness in aiding the detection of cancer, however, accuracy and precision are crucial. They have been selected as the measures to assess the suggested model. The accuracy rate, which is quantified by a mathematical formula, is the percentage of properly identified samples among all the samples. The accuracy rate is defined as the ratio of correctly identified positive samples to the total number of positive samples, as stated in the equation in Figure 4.2:

$$\mathrm{ACC} = \frac{(\mathrm{TP} + \mathrm{TN})}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}.$$

Figure 4.2: Accuracy Measurement

The formulas mentioned above use the following terms: TP for the proportion of positively predicted positive samples that actually were positive, TN for the proportion of accurately anticipated negative samples, FP for the proportion of falsely anticipated positive samples from negative samples, and FN stands for the quantity of positive samples that were mistakenly classified as negative.

4.3 Result

Each scaled 200x200 pixel image was split into 64 patches where each patch has resolution of 25x25. The patches were converted to tokens with positional embedding. For leukemia cells, we have labeled them as '1' and for healthy cells labeled them as '0'. The labeled tokens were feed into Vit model to perform classification.

The batch size is taken as 32. We have set the train test split to a 9:1 ratio to train the dataset. As the training time complexity is very high, we have set epochs as 17.

We have obtained a Training accuracy of **83.8%**, validation accuracy of 78.9% within 17 epochs. and training loss of 0.39.

By using a finely tuned database the accuracy could be improved. As the dataset is around 103GB, having advanced hardware can give drastic improvement and better training time. The image dataset was around 10 GB and the training time taken was 21 hours. The below graphs shows Epoch vs accuracy and Epocs vs loss in Figure 4.2 where the epoch is 17.

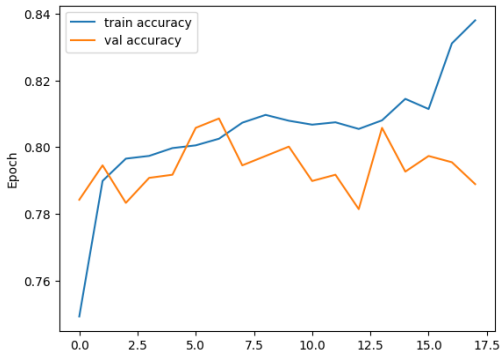


Figure 4.3: Accuracy Graph

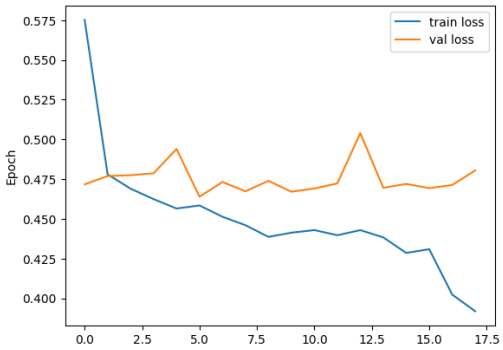


Figure 4.4: Loss Graph

Chapter 5

Conclusion

5.1 Future Work

Using a larger and more complex dataset could help to further extend the capabilities of our work. Normal image classification involves analyzing images in the visible spectrum, which includes the colors that are visible to the human eye. In contrast, hyper spectral images capture the entire spectrum of light reflected or emitted by an object or scene, and they can contain a wealth of information about the materials and substances present in the scene. By analyzing these images under a vision transformer, it is possible to study cancer cells and their chemical composition, which can be useful for detecting cancer and monitoring treatment. While hyper spectral image classification can be more challenging due to the larger and more complex dataset, the additional information in these images can also make them a powerful tool for identifying specific materials and substances.

5.2 Conclusion

In order to help doctors make accurate diagnoses in practical situations, this article describes an assessment and diagnosis for acute lymphocytic leukemia that uses the ViT model to separate malignant cells from healthy ones. The ISBI 2019 dataset was utilized in the study. Data scaling was done to overcome over-fitting and to train the model. ViT has exhibited improved functionality that processes visual data via self-attention between image patches. While showcasing the advantages of deep learning for blood cancer detection, we also highlighted the drawbacks and looked for a more effective strategy. The accuracy of this model is 83.8% . The results showed that this approach is far more flexible, and time-saving and could contribute to the world of medical diagnosis in real-time. Other researchers can use the info and knowledge from this paper and reach new limits. The code base for ViT model implemented in this paper is as follows: https://github.com/Sifat-ul-Amin/Vision-Transformer-ViT.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] N. Ahmed, A. Yigit, Z. Isik, and A. Alpkocak, "Identification of leukemia subtypes from microscopic images using convolutional neural network," *Diagnostics*, vol. 9, no. 3, p. 104, 2019.

[3] L. Boldú, A. Merino, S. Alférez, A. Molina, A. Acevedo, and J. Rodellar, "Automatic recognition of different types of acute leukaemia in peripheral blood by image analysis," *Journal of Clinical Pathology*, vol. 72, no. 11, pp. 755–761, 2019.

[4] A. Lavric and P. Valentin, "Keratodetect: Keratoconus detection algorithm using convolutional neural networks," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.

[5] L. Ma, C. Ma, Y. Liu, and X. Wang, "Thyroid diagnosis from spect images using convolutional neural network with optimization," *Computational intelligence and neuroscience*, vol. 2019, 2019.

[6] S. K. Asare, F. You, and O. T. Nartey, "A semisupervised learning scheme with self-paced learning for classifying breast cancer histopathological images," *Computational Intelligence and Neuroscience*, vol. 2020, 2020.

[7] N. Bibi, M. Sikandar, I. Ud Din, A. Almogren, and S. Ali, "Iomt-based automated detection and classification of leukemia using deep learning," *Journal of healthcare engineering*, vol. 2020, 2020.

[8] A. Bodzas, P. Kodytek, and J. Zidek, "Automated detection of acute lymphoblastic leukemia from microscopic images based on human visual perception," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 1005, 2020.

[9] A. Brodzicki, J. Jaworek-Korjakowska, P. Kleczek, M. Garland, and M. Bogyo, "Pre-trained deep convolutional neural network for clostridioides difficile bacteria cytotoxicity classification based on fluorescence images," *Sensors*, vol. 20, no. 23, p. 6713, 2020.

[10] M. I. Daoud, S. Abdel-Rahman, T. M. Bdair, M. S. Al-Najar, F. H. Al-Hawari, and R. Alazrai, "Breast tumor classification in ultrasound images using combined deep and handcrafted features," *Sensors*, vol. 20, no. 23, p. 6838, 2020.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] A. Gupta, R. Duggal, S. Gehlot, R. Gupta, A. Mangal, L. Kumar, N. Thakkar, and D. Satpathy, "Gcti-sn: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images," *Medical Image Analysis*, vol. 65, p. 101 788, 2020.

[13] P. H. Kasani, S.-W. Park, and J.-W. Jang, "An aggregated-based deep learning method for leukemic b-lymphoblast classification," *Diagnostics*, vol. 10, no. 12, p. 1064, 2020.

[14] Z. Lei, C. Zhou, S. Chen, Y. Huang, and X. Liu, "A sparse transformer-based approach for image captioning," *IEEE Access*, vol. 8, pp. 213 437–213 446, 2020.

[15] A.-A. Nahid, N. Sikder, A. K. Bairagi, M. A. Razzaque, M. Masud, A. Z. Kouzani, and M. P. Mahmud, "A novel method to identify pneumonia through analyzing chest radiographs employing a multichannel convolutional neural network," *Sensors*, vol. 20, no. 12, p. 3482, 2020.

[16] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Explainable covid-19 detection using chest ct scans and deep learning," *Sensors*, vol. 21, no. 2, p. 455, 2021.

[17] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.

[18] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for diagnosis of acute lymphoblastic leukemia based on vit-cnn ensemble model," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[19] R. Yang, Y. Du, X. Weng, Z. Chen, S. Wang, and X. Liu, "Automatic recognition of bladder tumours using deep learning technology and its clinical application," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 17, no. 2, e2194, 2021.

[20] W. Zhao, F. Chen, H. Huang, D. Li, and W. Cheng, "A new steel defect detection algorithm based on deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[21] N. Alanezi, E. Abdalhabib, A. Alfayez, D. Alsalman, F. Alanezi, S. Al-Rayes, S. Alyousef, H. AlNujaidi, A. K. Al-Saif, R. Attar, *et al.*, "Knowledge and awareness of leukaemia and its risks among the population of saudi arabia," *Informatics in Medicine Unlocked*, p. 100 971, 2022.

[22] N. Alanezi, E. Abdalhabib, A. Alfayez, D. Alsalman, F. Alanezi, S. Al-Rayes, S. Alyousef, H. AlNujaidi, A. K. Al-Saif, R. Attar, D. Aljabri, S. Al-Mubarak, M. M. Al-Juwair, L. Saraireh, N. Alenazi, and T. M. Alanzi, "Knowledge and awareness of leukaemia and its risks among the population of saudi arabia," *Informatics in Medicine Unlocked*, vol. 31, p. 100 971, 2022, ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2022.100971. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914822001162.

[23] Y. Lee and P. Kang, "Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder," *IEEE Access*, vol. 10, pp. 46 717–46 724, 2022.

[24] B. Li, J. Liang, and J. Han, "Variable-rate deep image compression with vision transformers," *IEEE Access*, vol. 10, pp. 50 323–50 334, 2022.

[25] S. NAHZAT, F. BOZKURT, and M. YAĞANOĞLU, "White blood cell classification using convolutional neural network," *Journal of Science, Technology and Engineering Research*, vol. 3, no. 1, pp. 32–41, 2022.