**Capstone Project-5: Leukemia Cancer Detection using Image Classification**

PROJECT REPORT

**Data Science with Python Programming**

**INDUSTRIAL PROJECT BASED LEARNING**



**Department of Computer Science and Engineering**

**Accredited by NBA**

**Geethanjali College of Engineering and Technology**

**(UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi)

Cheeryal (V), Keesara (M), Medchal.Dist.-501 301.

**By Team -4**

**Team Details :**

Shivanoor Vignesh 21R11A0597

G. Sandhya Rani 21R11A0569

R. Dhathri 21R11A0545

Hasini 21R11A0518

A .Pooja 21R11A0501

PSBN Sriya 21R11A0595

N Sai Charan 21R11A0588

MD. Akbar Ali 21R11A0586

# ABSTRACT

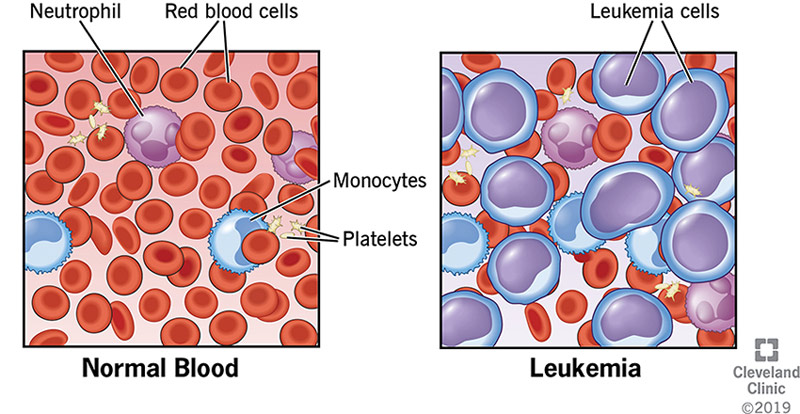
# Acute lymphoblastic leukemia (ALL) is a rare but serious form of blood cancer caused by the overproduction of abnormal lymphocytes in the bone marrow. Early and accurate detection is crucial for effective treatment and improved survival rates, especially in adults where the prognosis is often poor if diagnosed at a later stage. Current diagnostic methods rely on manual microscopic analysis of blood smear images, which is time-consuming, subjective, and prone to errors, particularly when distinguishing between normal and malignant cells with similar morphologies. This project proposes an intelligent deep learning approach to automate the screening of white blood cells for the presence of leukemia using microscopic blood smear images. Convolutional neural networks (CNNs), including architectures such as ResNet and VGG, are employed to classify images of blood cells as either normal or leukemic. The project involves preprocessing the image data through augmentation techniques, training and evaluating the performance of various CNN models, and identifying the most accurate classifier.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Chapter Name** | **Page No** |
|  | **ABSTRACT** | 2 |
| **1** | **INTRODUCTION** | 4 |
| **2** | **PROBLEM STATEMENT** | 6 |
| **3** | **OBJECTIVES** | 7 |
| **4** | **LITERATURE SURVEY** | 8 |
| **5** | **METHODOLOGY** | 9 |
|  | 5.1 Data Source | 9 |
|  | 5.2.1 importing the image dataset into python | 14 |
|  | 5.2.2 pre-processing the images using data augmentation | 14 |
|  | 5.2.3 Exploratory Data Analysis | 14 |
|  | 5.2.4 using TensorFlow’s map function | 16 |
| **6** | **ALGORITHM** | 17 |
|  | 6.1 using CNN Model and MobileNet | 17 |
|  | 6.2 using VGG16 Model | 18 |
| **7** | **Final Implementation** | 20-25 |
| **8** | **Conclusion** | 26 |

# INTRODUCTION

Leukemia, a devastating form of cancer affecting individuals of all ages, poses a significant global health threat due to its impact on white blood cells and the bone marrow. Early and accurate diagnosis is crucial for effective treatment and improved survival rates. Current manual methods of analyzing blood samples through microscopic images are slow and prone to inaccuracies, especially given the similarity in appearance between leukemic and normal cells. To address this challenge, automated and machine learning algorithms are being developed to enhance the detection of leukemia.



To overcome these challenges, this project introduces an advanced approach using artificial intelligence (AI) and deep learning. By employing powerful algorithms like convolutional neural .This study focuses on the development of an intelligent deep learning algorithm that utilizes microscopic blood smear images to differentiate between normal and leukemic cells. Leveraging convolutional neural networks (CNN), ResNet, or VGG architectures, this algorithm aims to improve the efficiency and accuracy of leukemia detection. By training these algorithms on extensive datasets of blood smear images, the proposed method offers a promising approach to early screening and diagnosis of acute lymphoblastic leukemia (ALL), a rare but potentially curable form of blood cancer prevalent in children. Early detection through intelligent screening methods holds the key to enhancing treatment outcomes and reducing mortality associated with leukemia.

# 

Current diagnostic methods for leukemia rely on manual microscopic analysis of blood samples, a time-consuming and error-prone process. To overcome these limitations, this study proposes an intelligent deep learning algorithm for automated leukemia detection. This algorithm will leverage convolutional neural networks (CNNs), specifically ResNet or VGG architectures, to analyze microscopic blood smear images. By training the algorithm on a vast dataset of labeled images (normal vs. leukemic cells), the system aims to achieve high accuracy and efficiency in differentiating between healthy and cancerous cells. This approach holds promise for early detection and diagnosis, particularly for Acute Lymphoblastic Leukemia (ALL), a childhood blood cancer with high curability rates when identified early. By implementing intelligent screening methods, this study proposes a significant step towards improved treatment outcomes and potentially reduced mortality rates for leukemia patients.

Leukemia is a cancer of the blood that affects white blood cells. There are several different types of leukemia, each classified by the type of white blood cell that is affected. Some common types of leukemia include:

* Acute lymphocytic leukemia (ALL)
* Acute myeloid leukemia (AML)
* Chronic lymphocytic leukemia (CLL)
* Chronic myeloid leukemia (CML)

# PROBLEM STATEMENT

# The problem statement revolves around developing an intelligent deep learning algorithm that can effectively differentiate between normal and leukemic cells in blood smear images to enable early and accurate detection of leukemia. By leveraging advanced machine learning techniques and extensive datasets, the goal is to enhance the efficiency and accuracy of leukemia prediction, ultimately improving treatment outcomes and survival rates for individuals affected by this deadly disease.

# To address these challenges, advanced machine learning algorithms, particularly convolutional neural networks (CNN), are being developed for the early and accurate detection of leukemia. These algorithms extract crucial information from blood cell images to differentiate between normal and abnormal cells efficiently. By training these algorithms on extensive datasets, researchers aim to improve the accuracy and efficiency of leukemia prediction, ultimately enhancing treatment outcomes and survival rates.

# Studies have shown promising results in using machine learning and deep learning techniques for the early detection of leukemia. Techniques such as feature selection, model optimization, and gene expression profiles are employed to achieve high levels of accuracy in classifying leukemia subtypes. The use of advanced algorithms like K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB) has shown significant accuracy in predicting leukemia, with SVM outperforming other algorithms with 90.0% accuracy.

# Objectives

* Import an extensive dataset of blood smear images for leukemia detection.
* Preprocess the images to enhance quality and remove noise using image processing techniques.
* Train a deep learning algorithm using ResNet for accurate cell differentiation between leukemic and healthy cells.
* Validate the algorithm's performance against gold standard methods for leukemia detection.
* Focus on early detection of acute lymphoblastic leukemia (ALL), a common type of cancer in children.
* Assess the clinical impact of the algorithm on treatment decisions and patient outcomes.
* Implement the algorithm to process blood smear images quickly and efficiently.
* Detect the stage of leukemia cancer based on the differentiation of cell types.
* Build a Flask API to enable easy integration of the algorithm into existing healthcare systems.

# 

# 4. Literature Survey

# Leukemia is a serious blood cancer requiring early and accurate diagnosis for successful treatment. Traditional microscopic analysis of blood smears is the current standard, but limitations such as slow processing time, subjectivity, and potential for errors necessitate the exploration of alternative methods.

# Deep Learning in Medical Image Analysis: (2017) suggest that CNNs can improve diagnosis and treatment planning in medical fields, including leukemia detection.

# Automated Blood Cell Analysis: (2016) explore methods for blood cell detection and classification, which could lead to accurate identification of leukemic cells.

# CNNs in Medical Imaging: (2017) discuss how CNNs can enhance disease diagnosis in medical imaging, potentially benefiting leukemia detection.

# Challenges in Leukemia Diagnosis: (2016) highlight challenges in manual examination for leukemia diagnosis, suggesting deep learning as a solution for rapid and accurate detection.

# Deep Learning Architectures: LeCun et al. (2015) review deep learning architectures like CNNs, ResNet, and VGG, offering insights for developing leukemia detection algorithms.

# Clinical Impact of AI-based Diagnosis: (2017) discuss how automated diagnostic systems can improve accuracy, reduce workload, and enhance patient outcomes.

# Integration of AI in Healthcare: (2019) address challenges and opportunities in integrating AI algorithms into healthcare workflows, which could improve leukemia diagnosis.

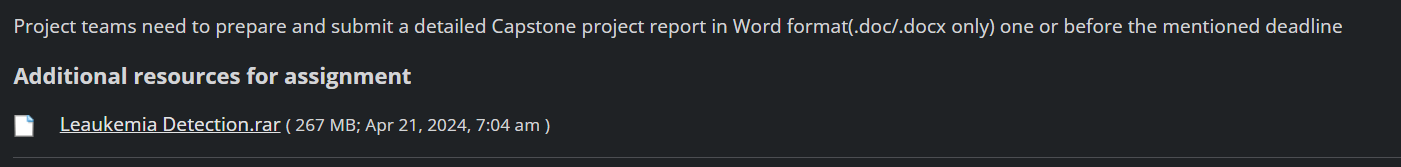
# METHODOLOGY

**5.1 Data Source**

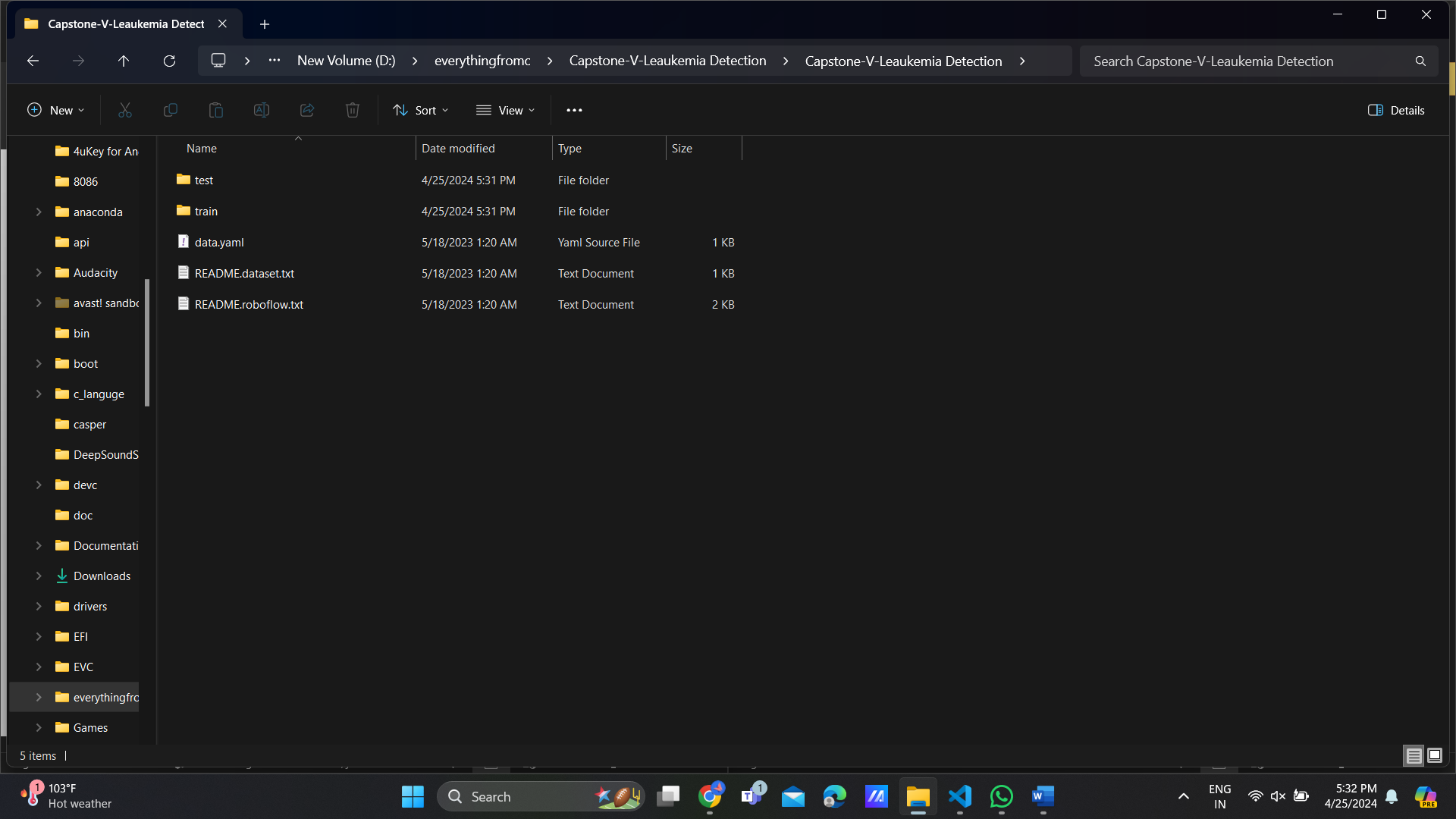
* **Brief description of the data source**

The data has been provided in the Sakai Learning Management System (LMS). The dataset consists of both training and testing data, with the training data containing images. The testing data includes both labels and an images folder.

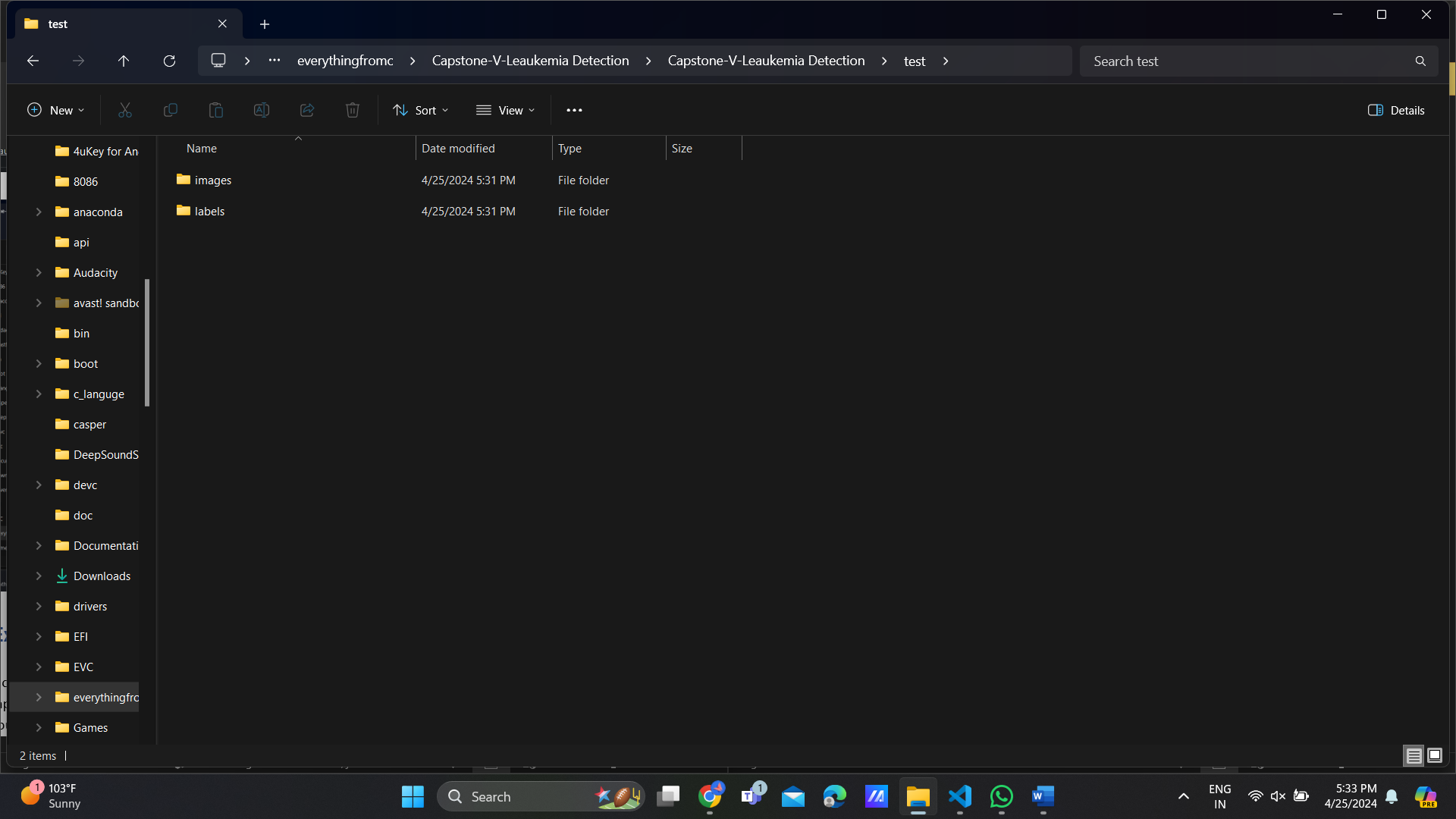
**Download the RAR file from the LMS:**

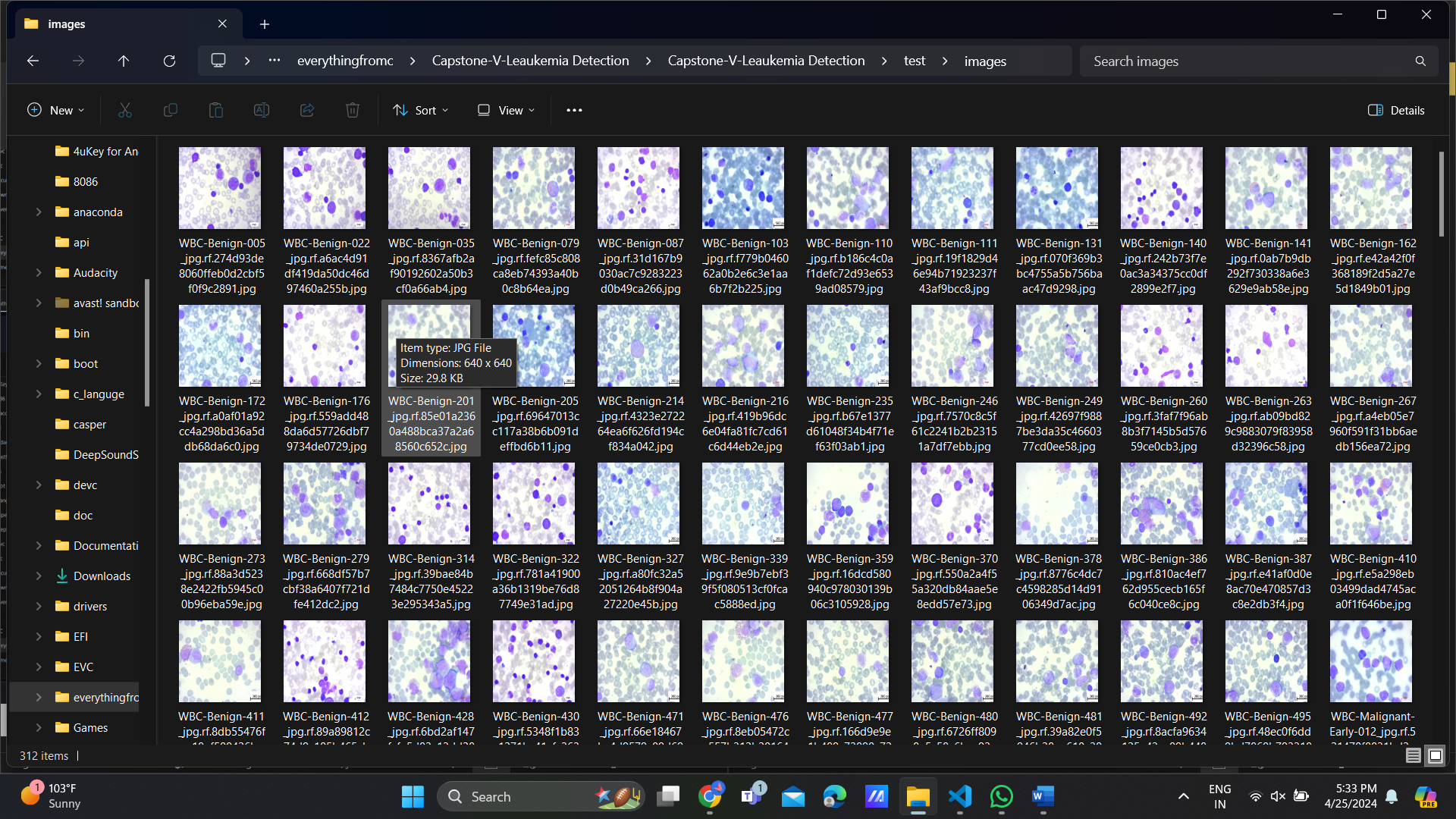


**Extract the folder from the RAR file and navigate to that folder:**

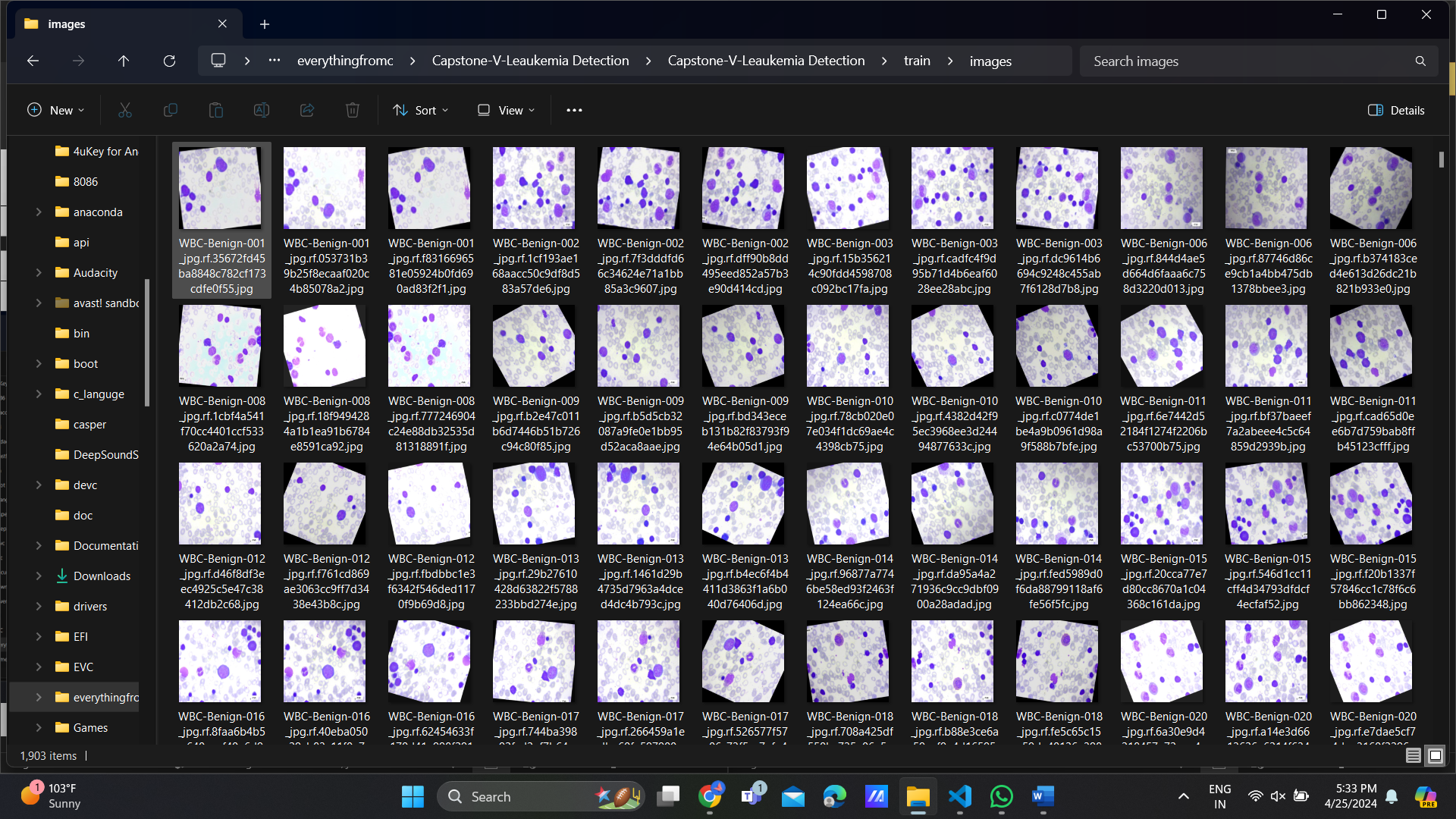


**Navigate to the test folder:**

   
  
**Open the Folder images**



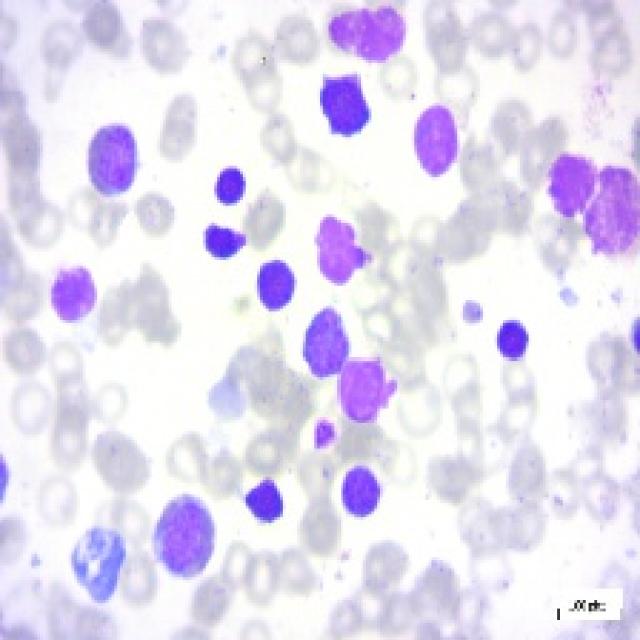
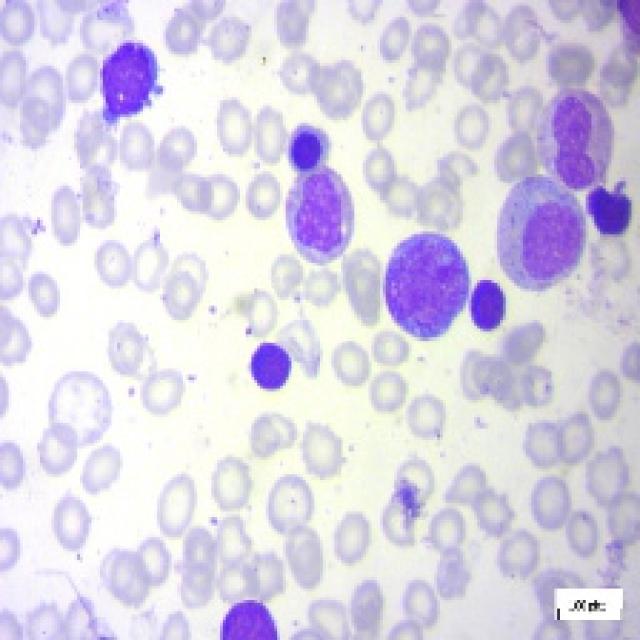
**NEXT Navigate to the train folder ->images**



**In leukemia ,cancer there are different stages include benign and malignant phases with the malignant stages further categorized as early , pre , pro**

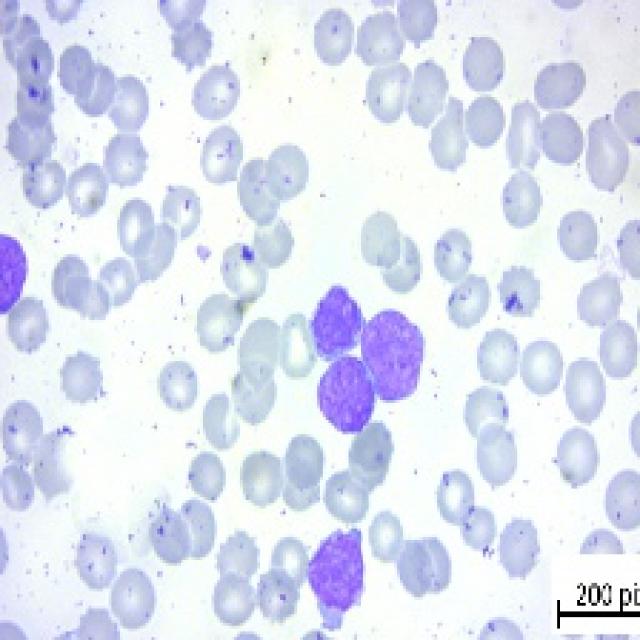
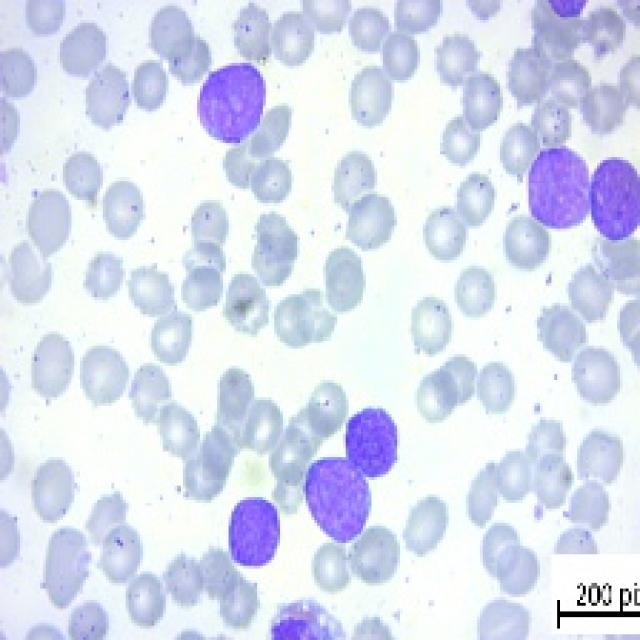
* The images illustrate various stages of leukemia

1.Bengin

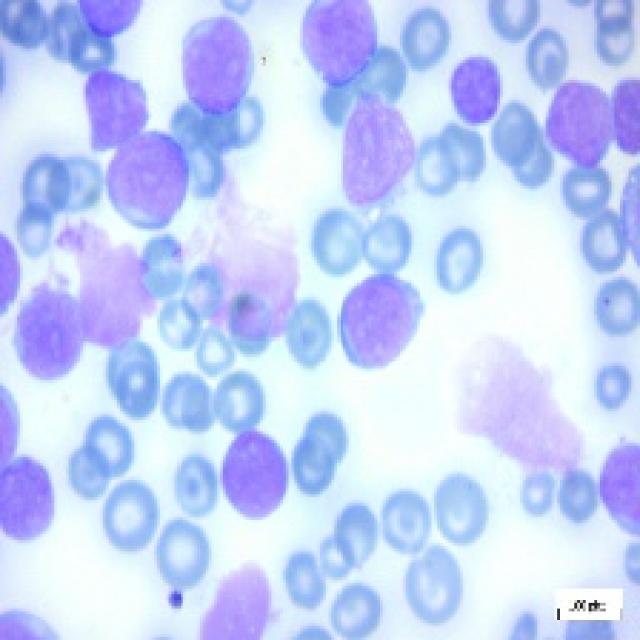
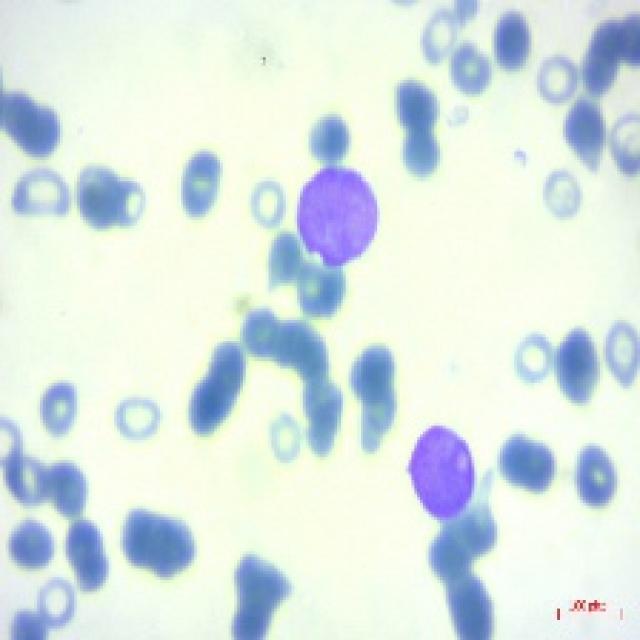
 

**2.maglignat**

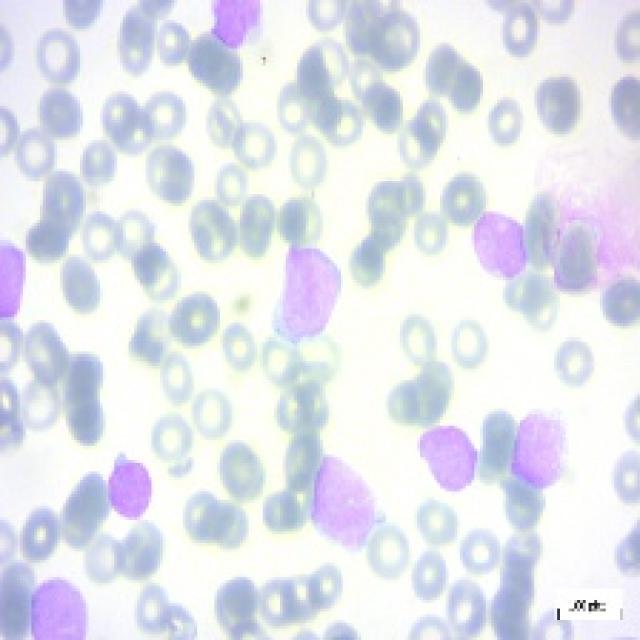
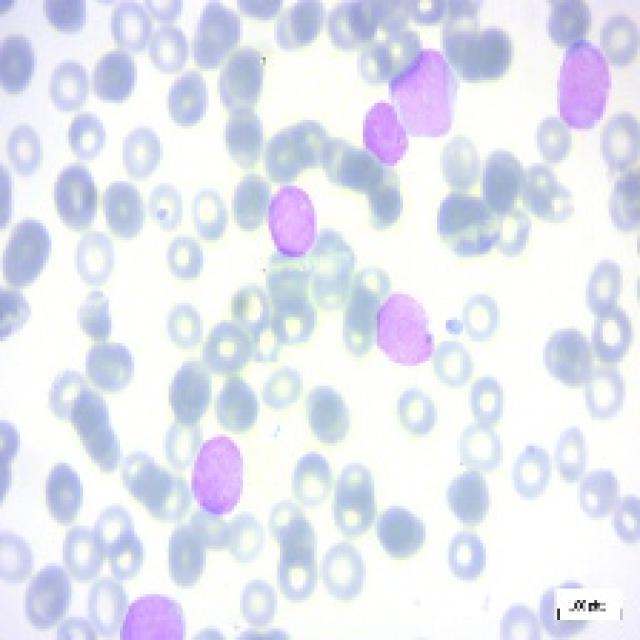
1. **early stage**

1. **maglignat pre**

1. **maglignat pro**



Leukemia encompasses a spectrum of conditions characterized by abnormal cell growth in the bone marrow. This progression occurs through several stages, each posing varying levels of severity and risk. The journey often begins with benign changes in blood cell development, which may not immediately manifest symptoms and require close monitoring. As abnormal cells multiply, they progress to a pre-leukemic stage, where the risk of developing leukemia increases, although the condition is not yet classified as cancer. The critical turning point comes with pro-leukemia, where the abnormal cells transition into full-fledged leukemia cells, although they may not have spread extensively. Early diagnosis and treatment are vital at this stage to effectively manage the disease.

* 1. **Methodology** 
     1. **Importing the image dataset into python:**

Utilize libraries such as OpenCV or PIL to import the dataset into Python. Ensure proper organization and import other necessary libraries like Keras and TensorFlow, including modules like Dense, Conv2D, MaxPooling2D, Flatten, and BatchNormalization. These libraries and modules are crucial for image processing and deep learning tasks in the project. Contains 7535 images classified in a Multi-Class Classification style for detecting leukemia cancer.

* + 1. **pre-processing the images using data Augmentation**
* The input images are in JPEG format and uses tf.image.decode\_jpeg to convert the raw data into tensors. The channels=3 parameter ensures these tensors represent RGB images with three color channels.
* Apply data augmentation techniques such as rotation, flipping, scaling, and cropping to increase the diversity of the dataset.
* All images are resized to a consistent size of [640,640] using tf.image.resize. This ensures compatibility with the deep learning model, as many models require images with uniform dimensions.
* Normalization to the 0-1 range by dividing by 255.0.
* Further normalization to the -1 to 1 range using multiplication by 2 and subtraction by 1. This is a common image processing technique that can improve model training and performance

# Exploratory Data Analysis

# After Loading the data from the sakai

# The data has training and testing set in which it has the classification of the

# Number of images in train: 6590

# Number of images in test: 312

# Number of labels in test: 312

# Annotations

# The annotations variable is a list that stores the bounding box annotations read from the text file. Each annotation is represented as a tuple containing five values:

# Class ID: An integer representing the class of the object.

# x\_center: The normalized x-coordinate of the center of the bounding box.

# y\_center: The normalized y-coordinate of the center of the bounding box.

# width: The normalized width of the bounding box.

# height: The normalized height of the bounding box.

# 

# From the above image says that

# The image shows a microscopic view of blood cells, specifically leukemia cancer cells.

# Green rectangles are drawn around certain cells, likely indicating areas of interest or specific features within the leukemia cells.

# The presence of leukemia cells can be inferred from the abnormal morphology and characteristics of the cells compared to normal blood cells.

# Different types of leukemia (e.g., acute myeloid leukemia, chronic lymphocytic leukemia) may have distinct features that require expert analysis for accurate diagnosis.

# Automated systems can assist pathologists in identifying and analyzing leukemia cells in such images, aiding in the diagnosis and treatment of leukemia

# Distribution of Labels/Categories

# 

The above bar graph shows the distribution of labels for leukemia cancer cells. The labels likely correspond to different types of leukemia. "benign" represent healthy cells,"malignant" represent cancerous cells.

# Using TensorFlow’s map Function

TensorFlow's map function is used to apply the load\_and\_preprocess\_from\_path\_label function to each element (image path and label pair) in the training and testing data.

# 

# The shape (196, 256, 3) indicates that the tensor is 3-dimensional. The first dimension (196) represents the height of the image, the second dimension (256) represents the width, and the third dimension (3) represents the number of color channels (RGB). The data type float32 specifies that the tensor elements are 32-bit floating-point numbers. This data type is commonly used for numerical computations in deep learning and machine learning due to its precision and efficiency

# 6. ALGORITHM

# 6.1 USING CNN Model and MobileNet Model

# 6.2 USING VGG16 model

# 6.1 Using CNN Model and MobileNet Model

# Giving the data of above data shape of(192,256,3) of three dimensional layer to added to reduce the spatial dimensions of the feature maps from the base model and obtain a single vector for each channel.

# Two dense layers are added on top of the flattened output. The first dense layer has 64 units and uses the ReLU activation function. The second dense layer has 5 units (equal to the number of classes) and uses the softmax activation function, which is suitable for multi-class classification tasks.

# 

The first layer is a MobileNetV2 layer with an output shape of (None, 7, 7, 12800) and approximately 2,257,984 parameters. The GlobalAveragePooling2D layer has an output shape of (None, 12800) and 0 parameters. The Flatten layer has the same output shape as the GlobalAveragePooling2D layer. The first Dense layer has an output shape of (None, 64) and 81,984 parameters. The second Dense layer (Dense\_1) has an output shape of (None, 5) and 325 parameters.

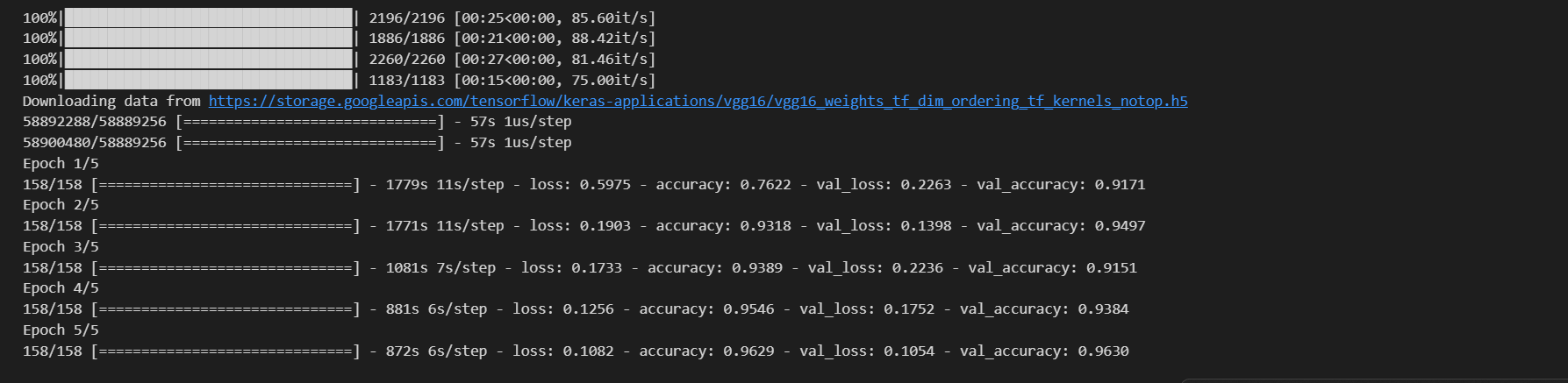
**ACCURACY**



MobileNet V2 has achieved an accuracy of 0.9716, which translates to 97% accuracy.

# 6.2 USING VGG16 model

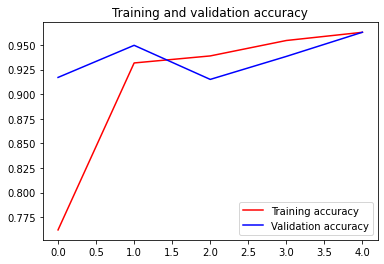
* Load images from different directories (categories), resize them to a fixed size (224x224), and store them along with their corresponding labels in lists X and y.
* Convert the lists of images and labels into NumPy arrays.
* Split the data into training and testing sets using train\_test\_split from scikit-learn.
* Normalize the pixel values of images to the range [0, 1].
* Convert categorical labels into one-hot encoded vectors using to\_categorical from TensorFlow.
* Load the VGG16 model pretrained on the ImageNet dataset, excluding the top classification layer.
* Freeze all layers in the base VGG16 model to prevent them from being trained.
* Add custom layers on top of the VGG16 base model to create the final classification model. This includes a GlobalAveragePooling2D layer followed by several dense layers with ReLU activation functions and a softmax output layer.
* Compile the model with the Adam optimizer and categorical crossentropy loss function.
* Train the compiled model on the training data for a specified number of epochs, using the validation data for evaluation.



 The model is being trained for 5 epochs.

 The training accuracy is increasing over time, while the validation loss is decreasing. This suggests that the model is learning to fit the training data well, but it is important to monitor the validation accuracy to avoid overfitting.

* The accuracy is over the value accuracy of 96%



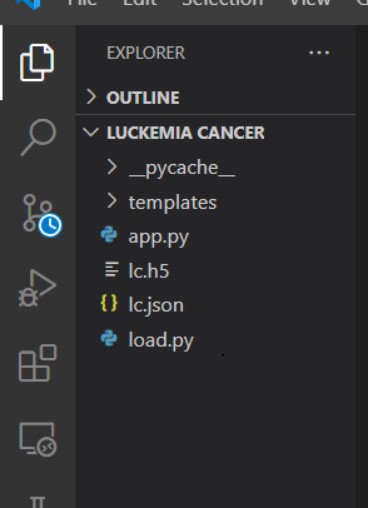
* Plot the training and validation accuracy over epochs using Matplotlib.
* The observation for this graph is that the training accuracy is higher than the validation accuracy.

**Final observation**

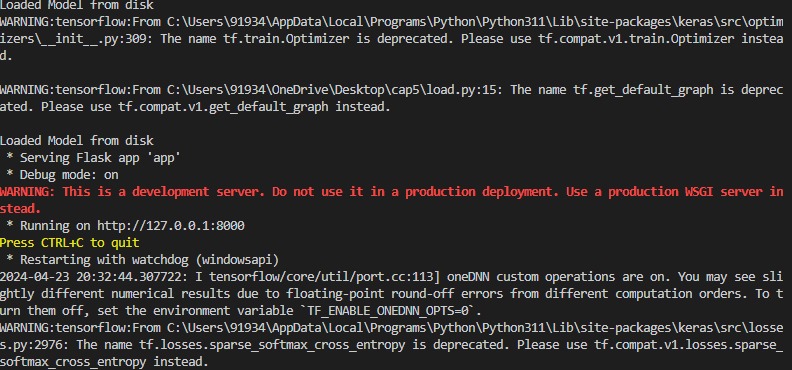
* Comparing the both models using cnn model and MobileNet model has the highest accuracy of 97% and while VGG16Model has 96% so final implementation has done in the CNN Model and MobileNet model

**6. FINAL IMPLEMENTATION**

**The file structure of the implementation of flask api for the leukemia cancer**



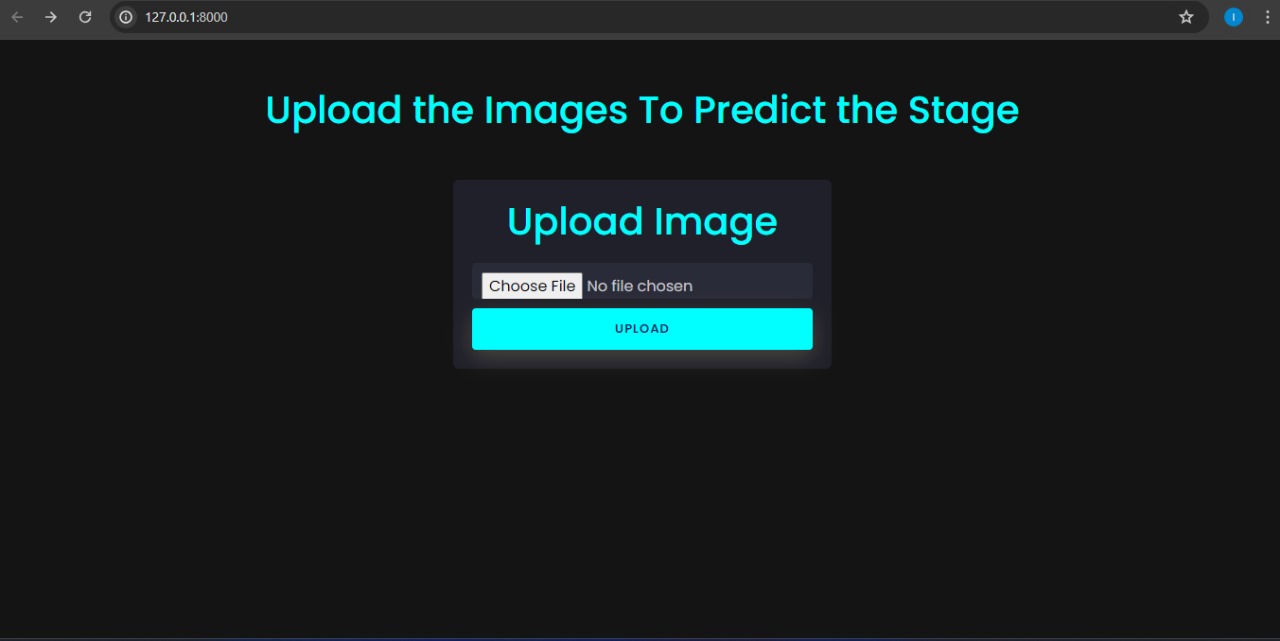
Click on falsk run to run the api as app.py in the terminal the display would be



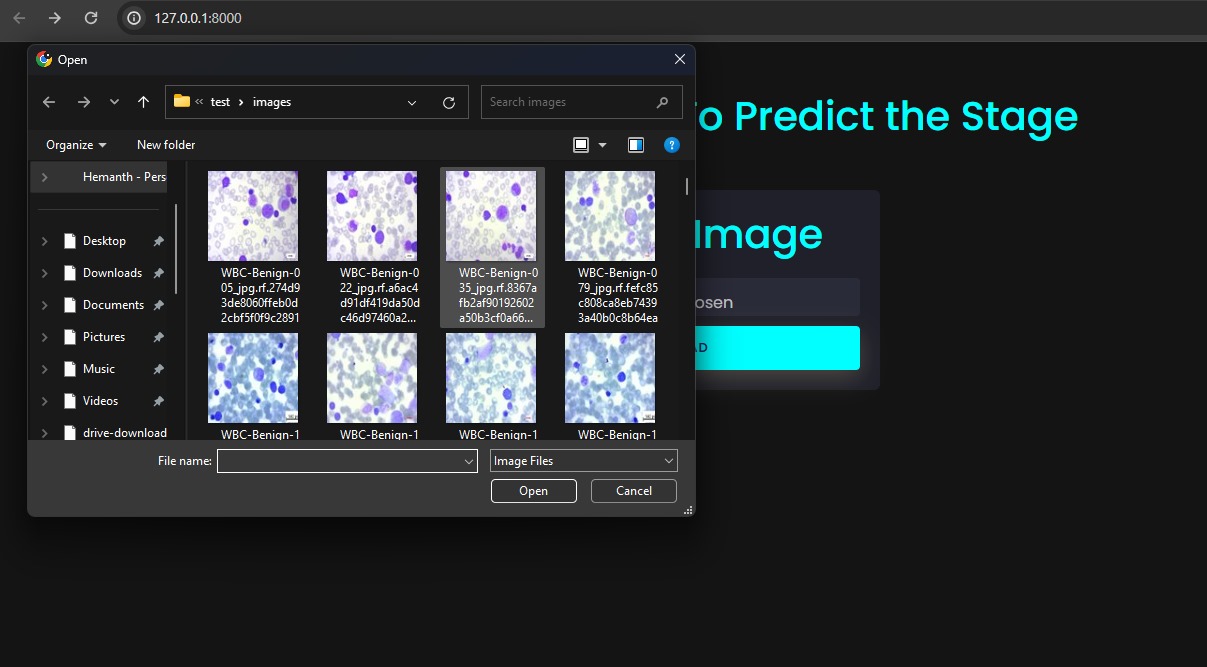
Click on the web page link of <http://127.0.0.1:8000> then web page is live now it predicts the stage of the leukemia cancer

By uploading the image of leukemia cancer

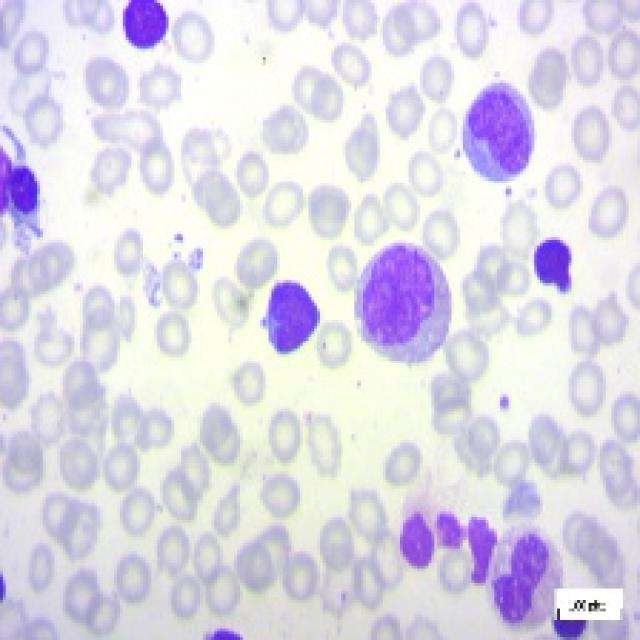
The website is :



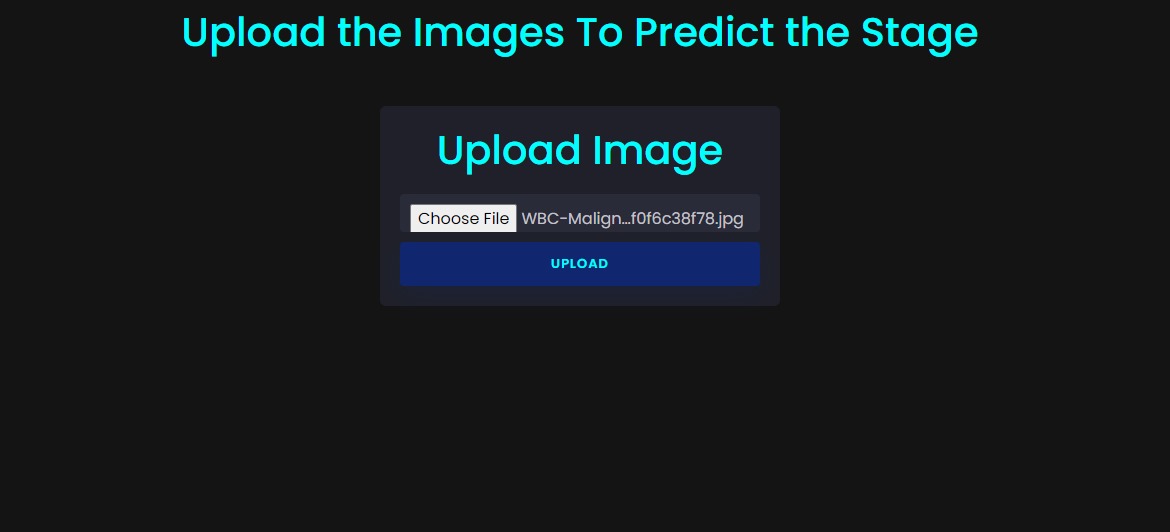
Click on the choose file for the predicting the stage of the leukemia cancer



The selected image for the prediction is the pictiure below which is selected from the test folder which data is provided



Click the image of the leukemia cancer then click on the upload button:

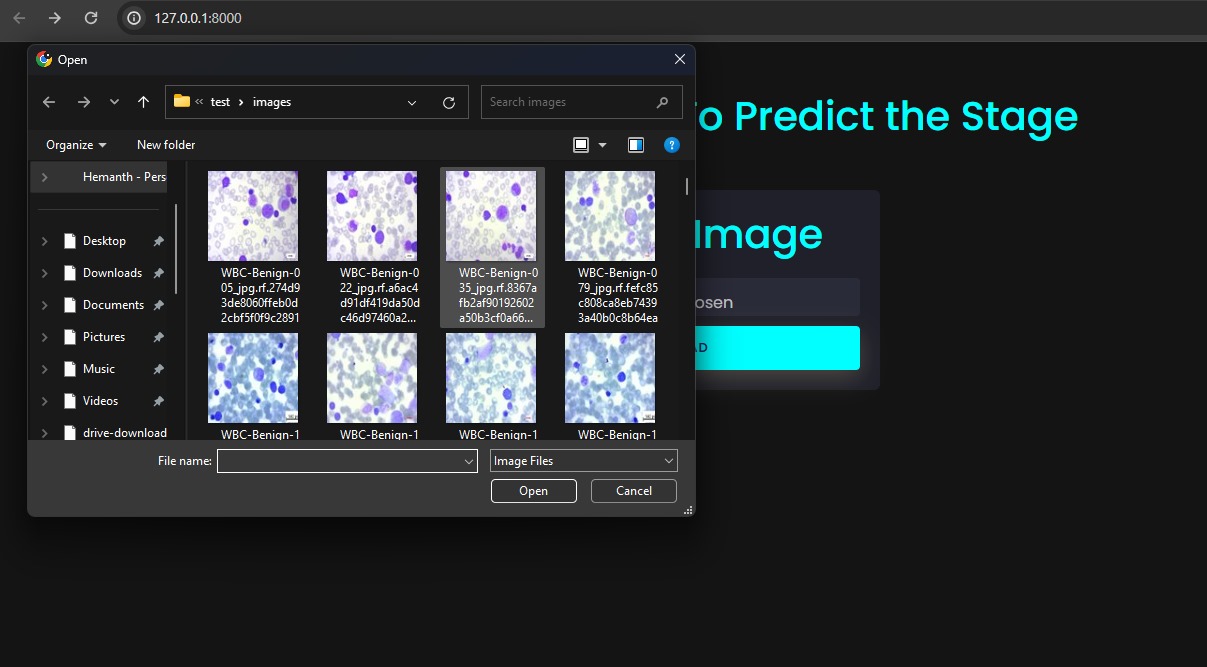


After clicking on the uploading button then in predicted label as the in result.html page

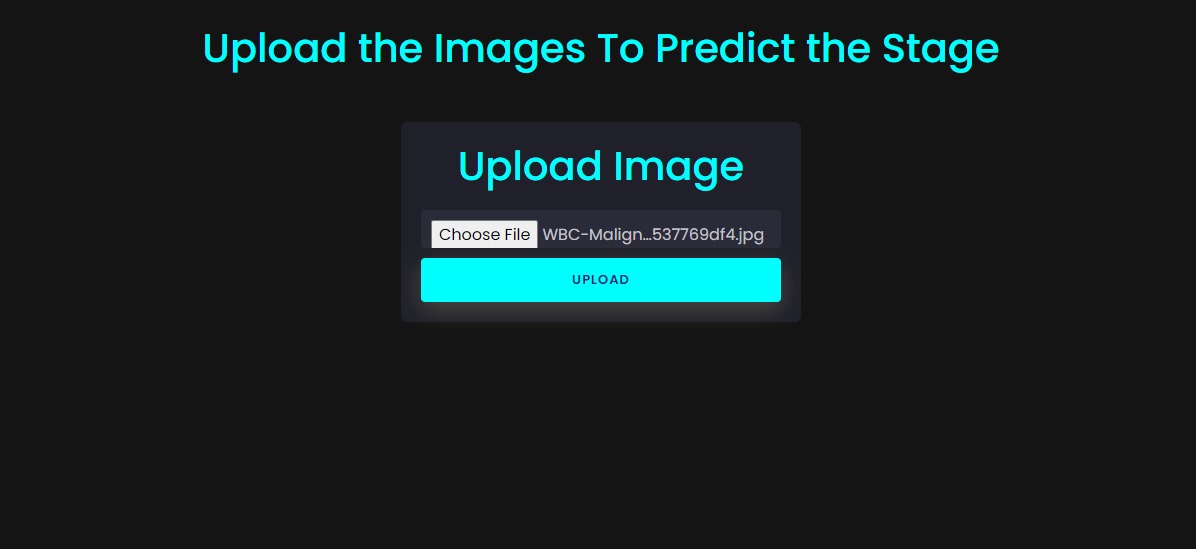


Predicting some more stages for the leukemia cancer :

Uploading the different image for the leukemia cancer



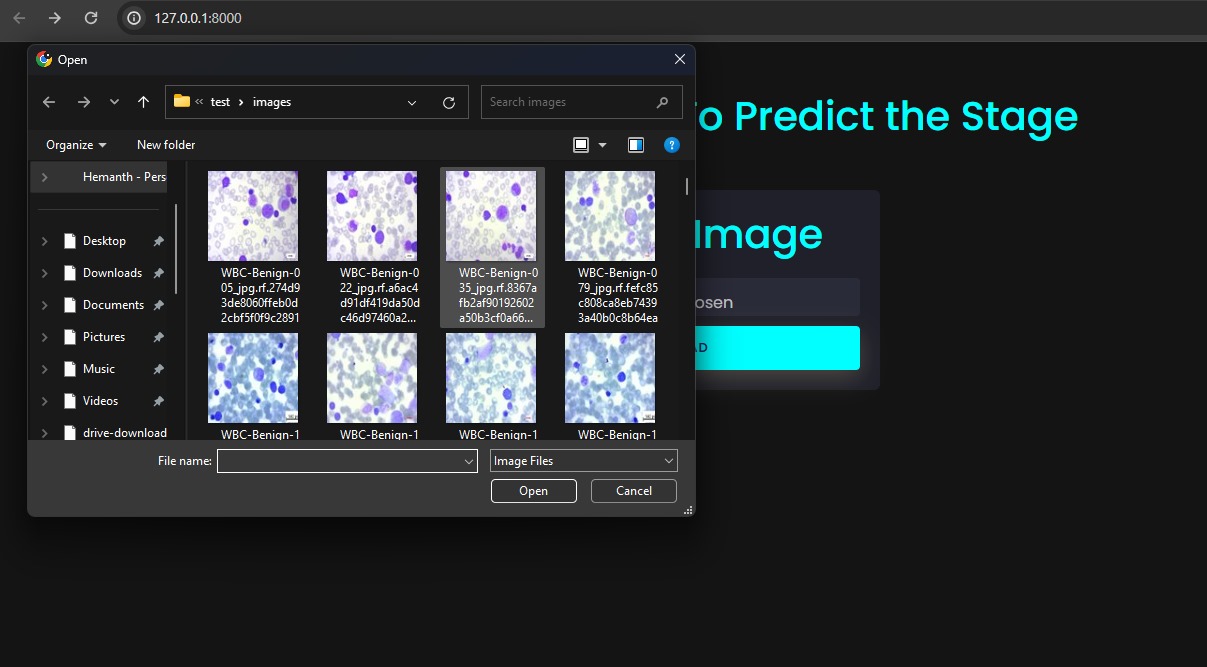
Selecting the picture and clicking on upload as the same as above



Prediction as the



Selecting another picture for the testing:



After uploading the picture it predicts as the



**7. Conclusion**

In this project, we developed an intelligent deep learning algorithm for the detection of leukemia cancer using image classification techniques. The algorithm utilizes Convolutional Neural Networks (CNNs), specifically the MobileNetV2 architecture, to analyze microscopic blood smear images and differentiate between normal and leukemic cells. Through extensive preprocessing, data augmentation, and model training, we achieved a high accuracy rate of 97% on the test dataset. This high accuracy demonstrates the effectiveness of our approach in automating the screening process for leukemia, potentially reducing the time and errors associated with manual analysis. Our algorithm has the potential to significantly impact the field of medical diagnostics by providing a faster, more accurate, and scalable solution for leukemia detection. Future work could involve expanding the dataset to include more diverse samples, optimizing the model further, and integrating it into clinical settings for real-time diagnosis and treatment planning.s