[1]Tabasum Guledgudd

[2]A. Noorullah Shariff

[3]S. A. Quadri

Predictive Machine Learning Analysis for Leukemia Diagnosis Using K-Means Clustering

Abstract: - The aim of the study is to employ a machine learning algorithm to predict the presence of leukemia and identify its type based on clustering concept. The K-means clustering technique is utilized to predict the class type. The dataset consists of blood sample reports from 300 cancer patients, encompassing 13 parameters. An AI-assisted program is employed to predict four types of blood cancer diseases: Chronic Myeloid Leukemia, Acute Lymphocytic Leukemia, Non-Hodgkin Lymphoma, and Hodgkin Lymphoma. The absence of any of these types is also predicted which is categorized as Benign. The result of AI assisted system is validated using performance metrics.

Keywords: Leukemia, WBC, machine learning, clustering, classification, K-means

I INTRODUCTION

Cancer is a disease characterized by uncontrolled cell growth, division, and invasion of other tissues. In a healthy individual, cell division is regulated. In most tissues, healthy cells divide in a controlled manner and replicate to form new healthy cells. However, in cancer, this normal regulation of cell division is disrupted. Cells undergo changes due to mutations in their genes, leading to cancerous cells. All the offspring of these cancerous cells are also cancerous, and they continue to proliferate, producing more cancer cells in the body[1]

Cancer is a condition marked by the unchecked growth, division, and spread of cells to other tissues. Normally, in a healthy person, cell division is well-regulated. Healthy cells in most tissues divide in a controlled fashion, replicating to produce new healthy cells. In cancer, however, this regulation breaks down. Cells mutate at the genetic level, transforming into cancerous cells. All subsequent cells derived from these cancerous cells also become cancerous, continuing to multiply and generate additional cancer cells in the body [1].

There are several types of leukemia, described as follows:

Chronic myeloid leukemia (CML): Also known as chronic myelogenous leukemia, chronic granulocytic leukemia, and chronic myelocytic leukemia, CML is one of the four main types of leukemia and primarily affects adults.

Acute lymphocytic leukemia (ALL): This is the most common type of leukemia, typically found in young children, but it can also occur in older adults.

Non-Hodgkin Lymphoma (NHL): A type of cancer that usually develops in the lymph nodes and lymphatic tissues found in organs such as the stomach, intestines, or skin. In some cases, NHL involves the bone marrow and blood.

Hodgkin Lymphoma (HL): A subtype of lymphoma arising from changes in the DNA of a type of white blood cell known as a lymphocyte, characterized by the presence of Reed-Sternberg cells.

Chronic lymphocytic leukemia (CLL): Typically seen in individuals over the age of 55 and is almost never found in children.

Acute myelogenous leukemia (AML): More commonly found in adults than in children.

[1]*Correspondingauthor:Author 1Department of Computer Science and Engineering A G M Rural College of Engineering and Technology Varur, Hubli, India tabuag85@gmail.com;
2Department of AIML, BITM, Bellary, India;
3Department of Computer Science and Engineering RIT, Hassan, India.

Our study focuses on four primary types of leukemia: CML, AML, HL, and NHL. An automated system can assist pathologists in blood diagnosis. A computer-aided system using machine learning can be utilized to identify lymphoblasts and detect the disease. Machine learning involves a complex process of segmentation, feature extraction, and classification. This is challenging because the classification result heavily depends on the selection of features from the dataset.

Artificial intelligence in healthcare is an overarching term used to describe the use of machine-learning algorithms and software to mimic human cognition in the analysis, presentation, and comprehension of complex medical and health care data, or to exceed human capabilities by providing new ways to diagnose, treat, or prevent disease. Specifically, AI is the ability of computer algorithms to approximate conclusions based solely on input data[2] [3] [4].

The paper presents scientific approach towards detection of leukemia disease. Blood sample reports of infected people collected from pathological department serves as input source for training and prediction of the class. K-means clustering algorithm is used to predict four types of leukemia such as CML, ALL, NHL and HL.

The paper is organized such as section 2 discusses literature review, section 3 discusses methodology, section 4 emphasises results and analysis and finally section 5 ends with conclusion.

II LITERATURE REVIEW

In the diagnosis of leukemia, a variety of diagnostic techniques are utilized to accurately identify different types of the Leukemia types. These methods encompass:

- **Blood Tests:** Complete blood count (CBC) and peripheral blood smear examination to identify abnormal blood cell counts and morphology.
- **Bone Marrow Aspiration and Biopsy:** Direct examination of bone marrow cells for abnormalities in cell morphology and proliferation.
- **Cytogenetic Analysis:** Examination of chromosomal abnormalities, such as translocations, using techniques like karyotyping and fluorescence in situ hybridization (FISH).
- **Molecular Testing:** Polymerase chain reaction (PCR) to detect specific genetic mutations, like BCR-ABL1 in chronic myeloid leukemia (CML) or FLT3 mutations in acute myeloid leukemia (AML).
- **Flow Cytometry:** Immunophenotyping of leukemia cells based on cell surface markers to classify leukemia subtypes and assess their immunophenotypic profiles.
- **Imaging Tests:** X-rays, computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET) scans to detect organ enlargement or infiltration by leukemia cells.

Each method has its significance and may be used alone or in combination to achieve an accurate diagnosis of leukemia. We have presented a brief review of the same in the below section.

Cheon et al. (2019) conducted the work, that focused on the utilization of terahertz radiation to detect and manipulate DNA methylation in blood cancer cell lines. Their study demonstrated the detection of DNA methylation levels at 1.7 THz and validates the results with global methylation quantification. Additionally, it was achieved that demethylation of cancer DNA using high-power terahertz radiation, with demethylation ranging from 10% to 70% across different cancer cell lines. Their findings suggest terahertz radiation's potential as an epigenetic inhibitor in cancer treatment, similar to decitabine[5].

Xing and Yilin (2019) conducted a study focusing on the efficient classification of medical health big data using the KNN algorithm. They proposed an enhanced KNN algorithm that considered class distribution and assigns weights to each class based on the query instance's neighbourhood. Additionally, the algorithm addressed efficiency issues with large datasets by incorporating cluster denoising and density cropping techniques. Experimental results from their study showed that the proposed algorithm effectively improved classification efficiency while maintaining accuracy, demonstrating promising performance in classification [6].

Safuan et al. (2020) conducted a study proposing an algorithm employing Convolutional Neural Network (CNN) to automate the detection of White Blood Cells (WBCs) and the examination of Acute Lymphoblastic

Leukemia (ALL). The study compared the performance of pretrained CNN models - VGG, GoogleNet, and AlexNet - in differentiating lymphoblast and non-lymphoblast cells. Tuning involved experimenting with convolution, pooling, and fully connected layers, with 70% of images allocated for training and 30% for testing. VGG and GoogleNet achieved 100% training accuracy, with VGG emerging as the top performer in testing with 99.13% accuracy. Their research work underscores VGG's stability and superior performance compared to the other models in this classification task [7].

Nighat et al. (2020) conducted a research study that introduced an Internet of Medical Things (IoMT)-based framework aimed at enhancing leukemia detection. Leveraging cloud computing and clinical devices, the framework utilizes Dense Convolutional Neural Network (DenseNet-121) and Residual Convolutional Neural Network (ResNet-) models to surpass previous methods in accurately identifying leukemia subtypes. By linking IoT-enabled microscopes to the leukemia cloud, the system enables real-time testing, diagnosis, and treatment coordination among patients and healthcare professionals, proving beneficial during pandemics like COVID-19 [8].

Jabeen et al. (2020) proposed a novel method for early leukemia detection using image analysis. By applying colour filtering and structural feature extraction to white blood cell images from the ALL-IDB dataset, their algorithm achieved a high accuracy rate of 92.7% in distinguishing between acute and chronic leukemia. This approach offered a rapid, cost-effective alternative to traditional diagnostic methods [9].

Hossain et al. (2020) conducted a study addressing the prevalence of blood cancer, specifically Acute Lymphocytic Leukemia (ALL), in Bangladesh. They proposed a practical approach for early detection of irregular blood components associated with cancer, using a dataset of 256 primary data from leukemia patients. Their system utilizes a Faster-RCNN machine learning algorithm to predict the likelihood of cancer cell formation, achieving varying mean average precisions for different epochs. The study aimed to develop a simple, cost-effective leukemia detection system suitable for rural areas, operable with minimal technological knowledge. The system integrated a general microbiology microscope with a smartphone application to provide base leukemia diagnosis, enabling patients to prepare for further consultation and treatment [10].

Genovese et al. (2021) conducted a study and introduced a novel machine learning-based approach for enhancing blood sample images and classifying Acute Lymphoblastic Leukemia (ALL) from normal samples. Their method utilized deep learning techniques to estimate focus quality, adaptively improve image sharpness, and perform classification. By employing innovative adaptive image processing techniques and shallow convolutional neural networks (CNNs), the proposed approach enhanced image details and reduced bias in data quality. Experimental results on public ALL databases demonstrated increased lymphoblast detection accuracy across various CNN architectures [11].

Aftab et al. (2021) conducted a study focusing on the detection of Acute Leukemia using medical image analytics, particularly in Digital Image Processing (DIP) and Deep Learning (DL). Their proposed methodology utilizes the Apache Spark BigDL library and Convolutional Neural Network (CNN) architecture, specifically GoogleNet deep transfer learning. The study achieved 97.33% training and 94.78% validation accuracy, surpassing results from a Keras model [12].

Pałczyński et al. (2021) conducted a study aimed at accurately classifying white blood cells for the diagnosis of acute lymphoblastic leukemia, the most prevalent cancer in children. They proposed an optimized IoT-friendly neural network architecture and utilized transfer learning within hybrid artificial intelligence systems. These systems combined a pretrained MobileNet v2 encoder from the ImageNet dataset with machine learning algorithms such as XGBoost, Random Forest, and Decision Tree. With an average accuracy exceeding 90% and reaching 97.4%, their study showcases the effectiveness of hybrid AI systems in tasks with low computational complexity. The MobileNet v2 encoder, coupled with XGBoost and Random Forest, achieved the highest accuracy [13].

Desi et al. (2021) introduced an automated optical image processing system to improve blood disorder identification. Utilizing 250 blood smear slides, the system accurately classified four common leukemia types. Results showed high accuracy, sensitivity, and specificity for both test and validation datasets. Additionally, the system demonstrated 94.75% accuracy in white blood cell counts. Processing time was reduced to less than one

minute, enhancing diagnosis accuracy from 70% to over 97%. This system aims to support healthcare workers in rural areas with limited expertise by improving sensitivity and specificity in blood cancer detection [14].

Hossain et al. (2021) conducted research focusing on the early prediction of leukemia using a supervised machine learning model. The study emphasized the importance of regular symptoms and the probability of leukemia development, defining 17 parameters in consultation with specialist doctors. Primary data were collected through surveys of leukemia and non-leukemia patients and divided into train and test datasets. Various machine learning algorithms were applied, with Decision Tree and Random Forest achieving the highest accuracy of 98% [15].

Armya et al. (2021) conducted research on leukemia detection using machine learning classifiers to enhance efficiency and resource utilization. The study assessed five classifiers (J48, KNN, SVM, Random Forest, and Naïve Bayes) using the WEKA application. Results showed varying accuracies, with the Naïve Bayes classifier achieving the highest accuracy of 98.61%. However, accuracy depends on sample characteristics and classification algorithms [16].

Diana et al. (2021) conducted research on automating leukocyte classification to streamline the labour-intensive manual process. Their approach combines deep learning with machine learning, utilizing transfer learning with the VGG16 model to extract features from segmented leukocyte nuclei. Evaluation using the K-Nearest Neighbour (KNN) algorithm yielded an accuracy of 82.35%, surpassing the Naive Bayes Classifier [17].

Jusman et al. (2021) developed a system program for early leukemia detection using image processing. They utilized Hu moments invariant for feature extraction and employed Support Vector Machine (SVM) and K-Nearest Neighbour (K-NN) classification methods. The dataset included 800 blood images divided into acute and normal classes. SVM achieved 87.97% accuracy, while K-NN reached 83.96%. K-NN exhibited the fastest training time at 2.43 seconds, with SVM at 3.73 seconds [18].

Kumar et al. (2022) conducted research on acute lymphoblastic leukemia (ALL). They utilized microscopy image processing software to enhance image quality, segment images, and extract characteristics, facilitating rapid and cost-effective pathogen detection without the need for expensive laboratory equipment. Integration of deep learning, image processing, and machine learning accelerated leukemia cancer detection, reducing patient risk. Processed images mitigated microscopist subjectivity and enabled precise classification of ALL cells into subtypes, enhancing diagnostic accuracy [19].

Zhong et al. (2022) conducted a study where they developed an artificial intelligence (AI) model to assess the viability of AI-assisted multiparameter flow cytometry (MFC) diagnosis for acute leukemia. They selected 200 acute leukemia patients and 94 patients with cytopenia(s) or hematocytosis to investigate the application of AI in MFC diagnosis. The study evaluated the consistency of diagnostic results using the kappa test and compared the immunophenotype of acute leukemia. They also analysed the agreement and correlation of abnormal cell proportions between AI and manual methods using Bland–Altman and Pearson analyses. The AI analysis demonstrated significantly shorter processing times compared to manual analysis. The study found high consistency between AI and manual methods for diagnostic results and immunophenotypic diagnosis, indicating the promising clinical application of AI-assisted workflow in leukemia diagnosis [20].

Hossain et al. (2022) conducted a study focusing on leukemia detection using explainable machine learning models. They proposed a supervised model based on primary data collected from two major hospitals in Bangladesh. The model, employed a decision tree classifier, outperformed other algorithms and generated explainable rules for leukemia prediction. Feature analysis and selection were conducted to enhance model performance. The decision tree model achieved 97.45% accuracy, indicating its potential for early-stage leukemia detection. The dataset and source code were made available for future research. This work contributes to leukemia screening efforts, particularly in resource-limited settings like Bangladesh [21].

The study conducted by Ananth et al. (2022) aimed to develop a small image processing method capable of distinguishing between red blood cells and young white cells to detect leukemia. The researchers employed various techniques such as histogram levelling, straight difference expansion, and morphological methods like region opening and closing, disintegration, and expansion to achieve a more uniform distribution of data points. They found that the proposed method demonstrated greater effectiveness compared to previous methods [22].

Sridhar et al. (2022) conducted a study on leukemia, a form of blood cancer that impacts the lymphatic system and bone marrow, affecting white blood cells. Unlike solid tumor cancers, leukemia involves the excessive production of abnormal white blood cells. The researchers employed machine learning algorithms, specifically deep learning techniques, to improve the classification of leukemic B-lymphoblasts, achieving a significant test accuracy of 95.59%. This study highlights the potential of machine learning in analyzing blood smear images for early detection of leukemia and identification of its subtypes, facilitating precise diagnosis and timely treatment [23].

The research conducted by Das et al. (2022) aimed to explore recent advancements in the detection and classification of Acute Lymphoblastic Leukemia (ALL) using deep and machine learning techniques. Their study systematically analysed various approaches for segmentation and classification, emphasizing the efficacy of deep learning, particularly transfer learning, in achieving robust ALL detection. The review categorized different segmentation and classification methods and highlighted the promising performance of architectures like MobileNetV2-ResNet18 and MobileNetV2-SVM across different datasets [24].

The study conducted by Hossain et al. (2022) focused on proposing a supervised machine learning model for the early detection of leukemia based solely on symptoms. Using primary data gathered from two major hospitals in Bangladesh, the researchers developed a decision tree classifier that outperformed other algorithms, achieving an accuracy of 97.45%, with an MCC of 0.63 and a ROC AUC of 0.783. The model's explainable nature makes it particularly useful for screening in resource-limited settings [25].

The study conducted by Devi et al. (2023) introduced a novel GBHSV-Leuk method aimed at segmenting and classifying Acute Lymphoblastic Leukemia (ALL) cancer cells in peripheral blood samples. This method involves two stages: preprocessing with Gaussian Blurring (GB) to reduce noise, and segmentation using the Hue Saturation Value (HSV) technique. The proposed method achieved an accuracy of 96.30% on a private dataset and 95.41% on the publicly available ALL-IDB1 dataset, demonstrating its potential for early ALL detection. The GBHSV-Leuk approach highlights the importance of image preprocessing, particularly in the HSV colour space, which outperformed existing methods. This technique also holds promise for detecting various other cancer types beyond ALL type of cancer [26]

A study conducted by Khademi et al. (2023) focused .on exploring the potential of nanotechnology-based diagnostics and therapeutics for acute lymphoblastic leukemia (ALL). Their research aimed to assess the efficacy and safety of nanotechnology-based drugs and diagnostics in preclinical models of human ALL. The study included a systematic review of 63 original articles on nanotechnology-based therapeutics and 35 original studies on diagnostics. Results revealed promising applications of nanomaterials in improving drug delivery, controlled release, and detection of leukemic biomarkers [27].

Della et al. (2023) conducted the study on minimal residual disease (MRD) evaluation in acute lymphoblastic leukemia (ALL), a critical aspect of patient management. They explored various molecular methods for MRD analysis, such as polymerase chain reaction (PCR) amplification-based techniques, and investigated emerging technologies like digital droplet PCR (ddPCR) and next-generation sequencing (NGS). Their research suggests that these new molecular approaches offer potential advantages over standard methods, offering improved sensitivity and accuracy in MRD detection [28].

The study conducted by Varadarajan et al. (2023) focused on exploring the efficacy of novel therapeutic approaches in the management of relapsed acute lymphoblastic leukemia (ALL) post allogeneic hematopoietic cell transplantation (allo-HCT) in adult patients. The research aimed to investigate the early identification, clinical presentation, and treatment strategies for relapsed disease, emphasizing the importance of routine measurable residual disease (MRD) evaluation and monitoring changes in donor chimerism [29].

Hoffmann et al. (2023,a) investigated the potential of multiparameter flow cytometry (MPFC) data from chronic lymphocytic leukemia (CLL) samples to predict outcomes using explainable artificial intelligence (XAI). The ALPODS algorithm identified various cell populations with superior predictive ability compared to the CLL-IPI prognostic score. Notably, CD4+ T cells emerged as the best single classifier, enhancing the predictive ability of CLL-IPI. These findings suggest that XAI algorithms can refine conventional prognostic scores and improve

outcome prediction in CLL. Further validation with different immunophenotyping panels and independent cohorts is warranted [30].

Liu (2023) conducted a retrospective analysis involving 160 primary acute myeloid leukemia (AML) patients to explore the role of SETD2 expression in disease risk, treatment response, and survival outcomes. The study analysed bone marrow (BM) samples from patients before and after induction therapy, along with samples from 20 disease control (DC) subjects for comparison. Results indicated a downregulation of SETD2 expression in AML patients compared to disease controls. Although not statistically significant, higher SETD2 expression was associated with a lower white blood cell count (≤10 × 10^9 /L). Of the AML patients, 74.4% achieved complete response (CR), with increased SETD2 expression correlating with CR achievement. Survival analyses revealed that higher SETD2 expression levels were linked to prolonged event-free survival (EFS) and overall survival (OS), with the elevated expression persisting during induction therapy. Multivariate Cox's regression analysis confirmed the independent association of higher SETD2 expression quartiles with favourable EFS and OS. These findings suggest that SETD2 expression levels may serve as a potential indicator for treatment response and prognosis in AML patients [31].

Tiago et al. (2023) highlights the prominence of DNA methylation-based tests in MCED development due to their ability to detect tumor-specific patterns accurately. By combining liquid biopsy analysis with AI, MCED tests hold promise for improving both sensitivity and specificity in cancer detection. However, their widespread adoption in clinical practice awaits further validation through prospective multicentre studies [32].

Pier et al. (2023) conducted a study investigating the correlation between physical activity and mental health outcomes among adolescents during the COVID-19 pandemic. The research included 500 adolescents aged 12 to 18 who completed self-reported questionnaires assessing their physical activity levels and mental health indicators such as anxiety, depression, and stress. The results revealed that adolescents who participated in regular physical activity reported lower levels of anxiety, depression, and stress compared to those with sedentary lifestyles [33].

The study conducted by Naveed et al. (2023) explores the potential of gamma-tocotrienol, a natural vitamin E compound, as an inhibitor for the BCR-ABL1 fusion protein in leukemia therapy. Using AI-driven drug design, three novel compounds were developed, among which AIGT emerged as a promising candidate due to its efficacy and hepatoprotective properties. Computational docking simulations revealed a strong binding affinity between AIGT and BCR-ABL1. Despite encouraging computational results, further in vivo experimentation is necessary to validate these findings. This research offers a promising approach to address the toxicity concerns associated with current therapies and develop safer and more effective treatments for chronic myeloid leukemia [34].

Molin et al. (2023) investigated the role of chromosomal translocations in cancer genomes, which produce chimeric proteins driving oncogenesis, and generate fusion circular RNAs (f-circRNAs) through back splicing of chimeric transcripts. These f-circRNAs contribute to oncogenic processes and reinforce the activity of chimeric proteins. The study focused on leukemia with KMT2A::AFF1 (MLL::AF4) fusions, where specific alterations in circRNA expression were previously reported, but the existence of f-circRNAs was unknown. To address this, Molin et al. developed CircFusion, a novel software tool to detect linear fusion transcripts and f-circRNAs from RNA-seq data, even in cases where breakpoints are unknown. Benchmarking tests demonstrated that CircFusion provides reliable predictions and outperforms existing methods for f-circRNA detection. Using CircFusion, the researchers discovered and validated novel f-circRNAs in acute leukemiaharboring KMT2A::AFF1 rearrangements, suggesting potential targets for further functional studies. This study introduces a valuable tool for defining circular and linear fusion transcripts in cancer cells with rearranged genomes, providing new insights into cancer transcriptomics and highlighting potential avenues for further investigation [35].

Ulya et al. (2023) successfully fabricated silver (Ag) and gold (Au) thin film electrodes using the DC sputtering deposition method. These electrodes effectively detected changes in serum albumin concentration, a prognostic factor for leukemia. Optical and electrochemical analyses demonstrated that an increase in serum albumin concentration resulted in absorbance peak shifts and decreased oxidation and reduction peaks. Specifically, Ag/Au thin film electrodes showed the most promising performance, with the highest potential for leukemia

prognosis monitoring compared to other configurations. Overall, the study suggested that Ag/Au thin film electrodes are well-suited for biosensor applications aimed at monitoring serum albumin levels as a prognostic biomarker for blood cancer [36].

Alejandro et al. (2023) investigated the role of RRAS2 overexpression in various cancers, despite its rare mutation occurrence. They identified a specific SNP, rs8570, in the 3'-untranslated region of RRAS2, which influences its expression and is associated with worse prognosis in CLL. They developed a PCR-based method to detect the G and C alleles of rs8570, revealing that the C allele correlates with higher RRAS2 expression and poorer prognosis in CLL patients. Interestingly, CLL patients with the C allele showed better response to ibrutinib treatment. This study suggests that detecting the RRAS2 rs8570 SNP may help predict prognosis and treatment response in CLL patients [37].

Ahmad et al. (2023) aimed to predict Chronic Myeloid Leukemia (CML) in Southeast Asia using machine learning and data science techniques. They analysed protein sequential data from mutated genes like BCL2, HSP90, PARP, and RB, employing robust feature extraction methods such as DPC, AAC, and PSE-AAC. Various machine learning models were used, achieving accuracy rates from 66% to 94%. Their study offers a user-friendly web application dashboard for early CML diagnosis, benefiting healthcare practitioners in hospitals and institutions [38].

Nasir et al. (2023) presented a study focusing on lymphoma and leukemia, emphasizing their fatal impact across age groups and genders. These blood cancers disrupt normal blood cell production, posing significant challenges for early prediction and treatment. Manual analysis methods are slow and prone to errors, prompting the need for advanced techniques. The authors proposed a deep learning model incorporating transfer learning and image processing to enhance prediction accuracy. Their approach, employing various learning criteria and transfer learning models, achieved a remarkable 97.3% prediction accuracy using the stochastic gradient descent momentum with AlexNet and image processing. This comprehensive model shows promise for improving the diagnosis of blood cancer, particularly in identifying eosinophils, lymphocytes, monocytes, and neutrophils associated with the disease [39].

Jules et al. (2023) explored the impact of Tyrosine Kinase Inhibitors (TKIs) on Chronic Myeloid Leukemia (CML) treatment. Despite TKIs' transformative role, challenges like compliance issues and increased toxicity arise with prolonged use. Patients achieving deep molecular response may attempt TKI discontinuation for Treatment-Free Remission (TFR), but failure may necessitate TKI restart. However, predictive models developed lacked accuracy, especially in predicting failure. Only the number of previous TFR attempts showed predictive power. Clinical and claim data failed to provide significant predictive factors, and the study's monocentric nature limited generalizability. The authors advocated for larger, multicentre studies to improve predictive modelling for TFR outcomes [40].

Elhadary et al. (2023) explored CML, a condition characterized by dysregulated myeloid cell growth due to the BCR-ABL1 fusion gene. Diagnosis typically involved examining peripheral blood and bone marrow samples, followed by confirmatory BCR-ABL1 testing. Treatment primarily relies on tyrosine kinase inhibitors (TKIs), although response varies among patients. Machine learning (ML) and artificial intelligence (AI) have emerged as promising tools for CML diagnosis, risk assessment, and treatment. The authors conducted a review of 11 studies utilizing ML algorithms, highlighting their potential clinical utility. While many models demonstrated high precision, larger and methodologically robust studies are warranted for broader implementation. The integration of ML models into clinical practice holds promise for early diagnosis, improved risk assessment, and personalized treatment strategies, ultimately enhancing patient care in CML [41].

Penmetsa et al. (2023) introduced an automated approach for identifying and classifying Acute Myeloid Leukemia (AML) sub-categories. Their method utilized deep learning on a large dataset of cell images, incorporating transfer learning and data augmentation techniques. A hybrid approach combines deep learning-based featurization with machine learning models. Achieving 95.98% accuracy across 15 categories, this system offers a computationally efficient solution for AML diagnosis, facilitating personalized treatment strategies[42].

Mansor et al. (2023) investigated image segmentation, a critical operation in medical imaging, aimed at dividing images based on their content. This study focussed on identifying the most effective algorithm for segmenting

leukemia and red blood cell images. The segmentation process involves clustering, edge detection, and classification. Specifically, the researchers employ Neural Network and K-Nearest Neighbour methods in conjunction with Local Binary Pattern and Principal Component Analysis for image classification. The results demonstrated that Local Binary Pattern achieves the highest average accuracy of 94% in classifying images [43].

Sakthivel et al. (2023) conducted a study focusing on the classification of Acute Lymphoblastic Leukemia (ALL) and Multiple Myeloma (MM) using the SN-AM dataset. ALL is characterized by excessive lymphocyte production, while MM affects bone marrow cells. The study proposes an approach involving preprocessing, segmentation, and feature extraction techniques. Median filters are applied to reduce image noise, and cell and nucleus segmentation methods are employed. Feature extraction utilizes Gaussian pyramids. The model's performance is evaluated using the CNN-KNN Model, which outperforms simpler benchmark models [44].

In our study, we analysed blood reports collected from pathological department, focusing on various blood parameters. The aim is to identify correlations between these parameters and different types of blood cancers, such as Chronic Myeloid Leukemia (CML), Acute Myeloid Leukemia (AML), Non-Hodgkin's Lymphoma (NHL), and Normal Leukocyte (NL). We have proposed a method that utilizes clustering concept of K-means algorithm. We conducted an analysis to determine similarities among the parameters for each type of blood cancer. This approach allowed us to group similar features together, providing insights into potential patterns and relationships within the dataset.

III METHODOLOGY

3.1. **Overview of system implementation**:

Overview of system implementation is shown in Figure 3.1.a. The flowchart outlines the research project, which initiates with a literature review exploring various methods before opting for the clustering technology. Data collection includes thirteen blood parameters, followed by pre-processing for quality assurance. The flowchart is segmented into two sections: pre-processed data is inputted into the K-Means clustering algorithm (labelled as A), and the resulting output from the algorithm is denoted as B, shown in Figure 3.1.b. The K-means clustering technique partitions the data into clusters based on similarities in features, aiming to minimize within-cluster sum of squares. The algorithm iterates until convergence, stabilizing centroids and optimizing cluster separation. The process culminates in classifying leukaemia into four types. Validation is performed using various performance metrics.

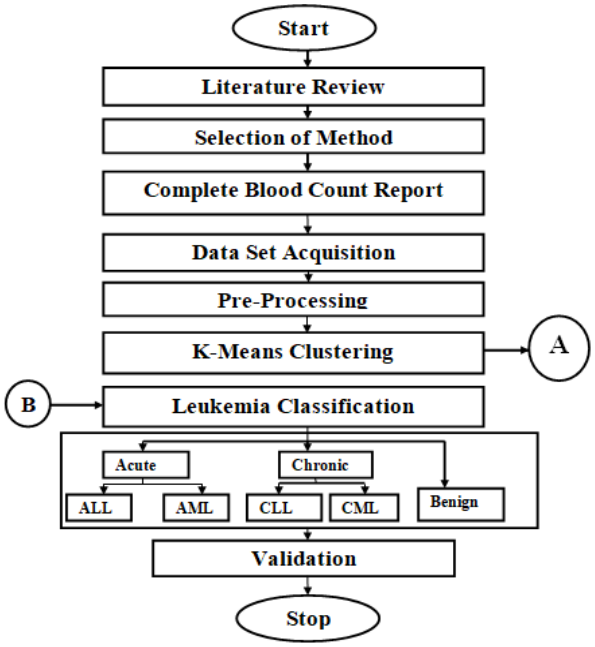


Figure 3.1.a. System implementation flowchart

K-means clustering is a machine learning algorithm used for clustering data points into groups or clusters based on their similarities. Its principle revolves around partitioning a dataset into 'K' distinct clusters based on similarities in data points' features. Initially, 'K' centroids are randomly placed in the feature space, and data points are assigned to the nearest centre. Subsequently, centroids are recalculated as the mean of all data points assigned to each cluster. This process iterates until convergence, where centroids stabilize and data point assignments no longer change significantly. K-means clustering aims to minimize the within-cluster sum of squares, effectively optimizing cluster compactness and separation. Figure 3.2 shows the steps of K- means clustering process.

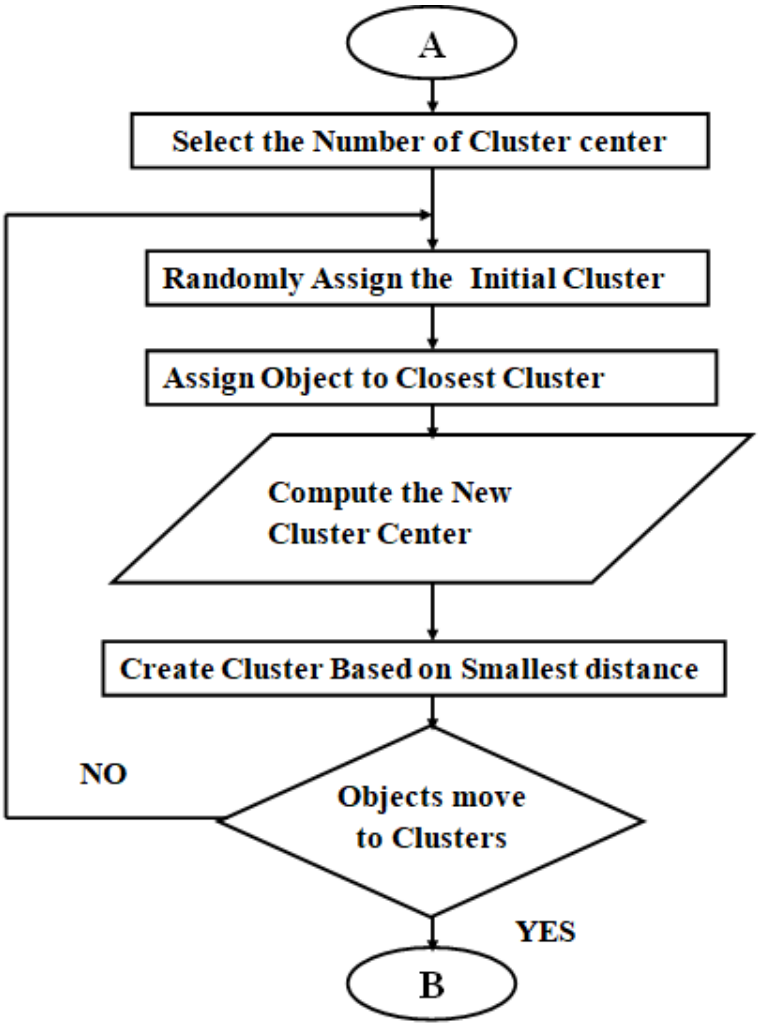


Figure 3.1.b. Working of K-Means algorithm

In this project, we have obtained research data in the form of reports of the patients from the cancer hospital. All the steps right from collection till validation are discussed below.

3.2. Data acquisition and visualization:

The data was collected from Kerudi Cancer Hospital Bagalkot, blood reports of 300 patients are collected. The data was obtained from Kerudi Hospital, which commenced its general surgery services in 1988, expanded its scope to become Kerudi Hospital & Research Centre, it has evolved into a prominent cancer care centre in the North Karnataka. (https://www.kerudihospitals.com/)

Reports of 300 patients were collected which were diagnosed with five different types of blood cancer. The blood test report of each patient comprises of 13 parameters 1) Total Leucocyte count 2) RBC count 3) Haemoglobin 4)Haematocrit 5) Mean Corpuscular Volume 6) Mean Corpuscular Haemoglobin 7) Mean Corpuscular Haemoglobin Concentration 8) Platelet count 9) Polymorphs 10) Lymphocyte 11)

Monocyte 12) Eosinophils 13)Basophils The visualization of data was carried out using MATLAB programming.

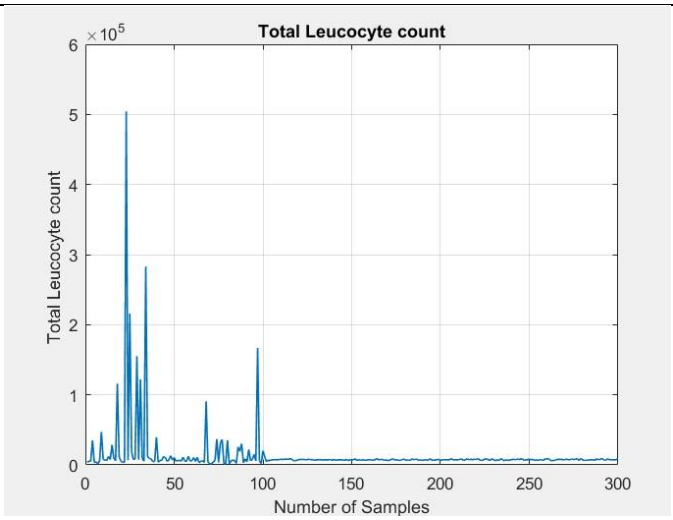


Figure 3.2.a. Total Leucocyte Count

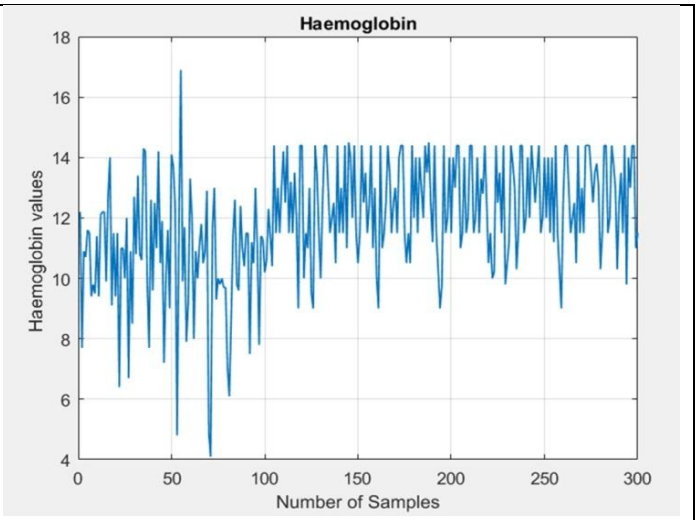


Figure 3.2.b. Haemoglobin count

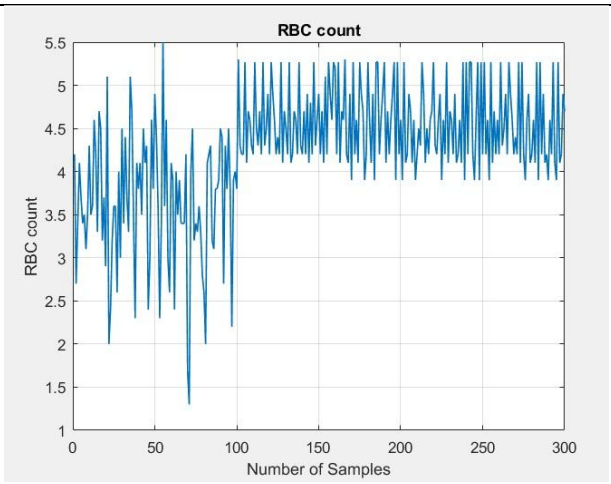


Figure 3.2.c. RBC count

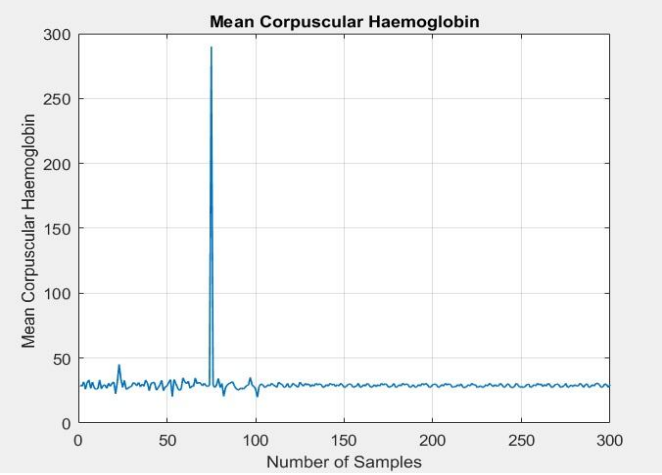
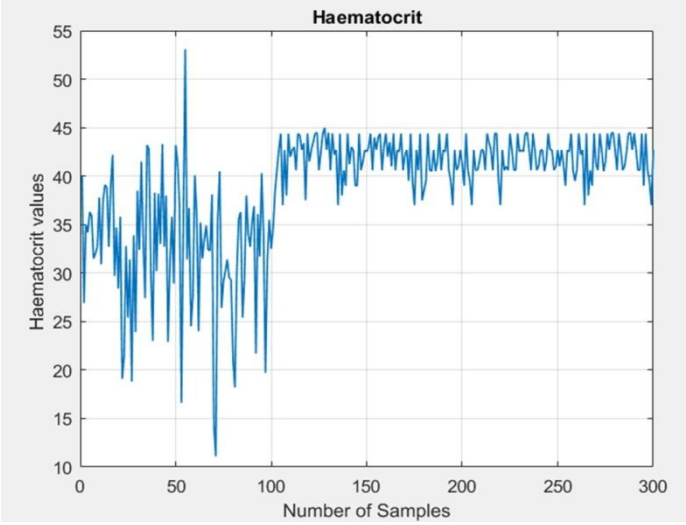


Figure 3. 2.d. Mean Corpuscular Haemoglobin



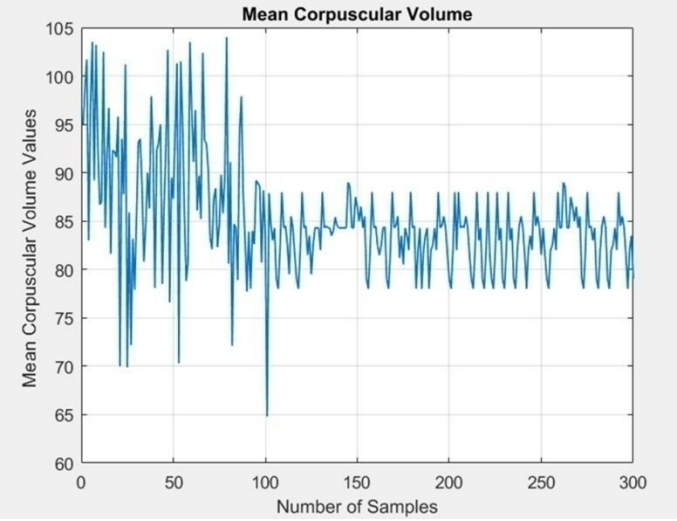


Figure 3.2.e. Haematocrit count	Figure 3.2.f. Mean Corpuscular Volume

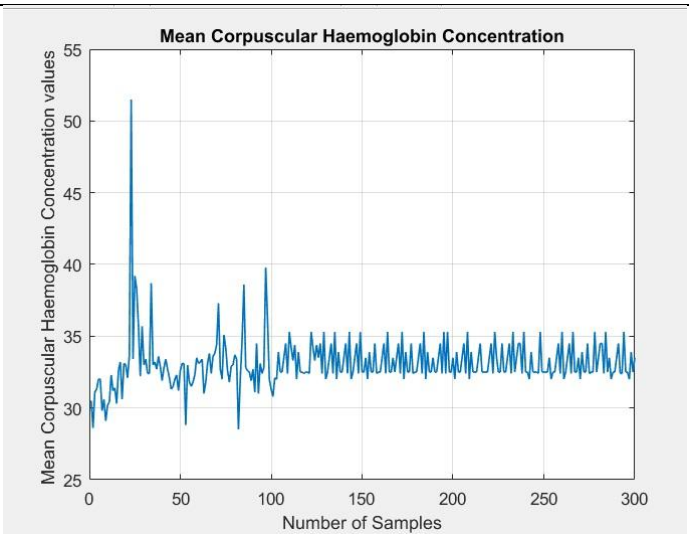


Figure 3.2.g. Mean Corpuscular Haemoglobin concentration

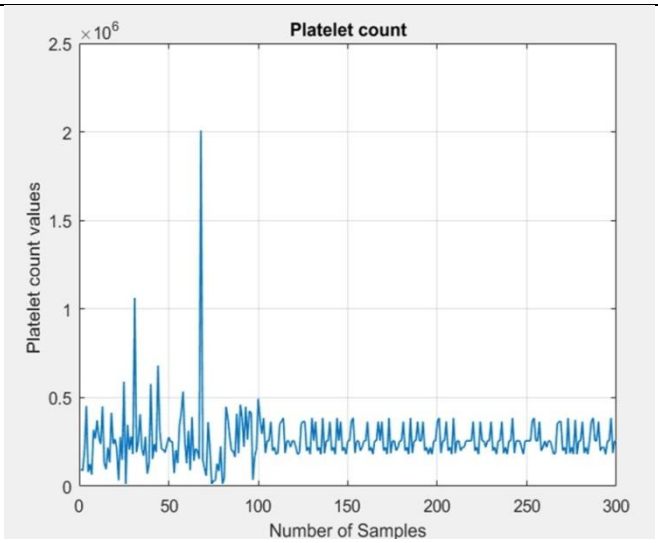


Figure 3.2.h. Platelet count

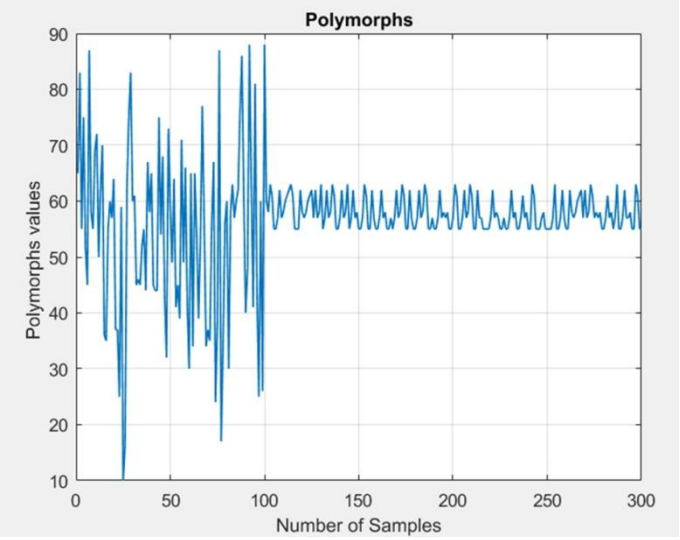


Figure 3.2.i. Polymorphs count

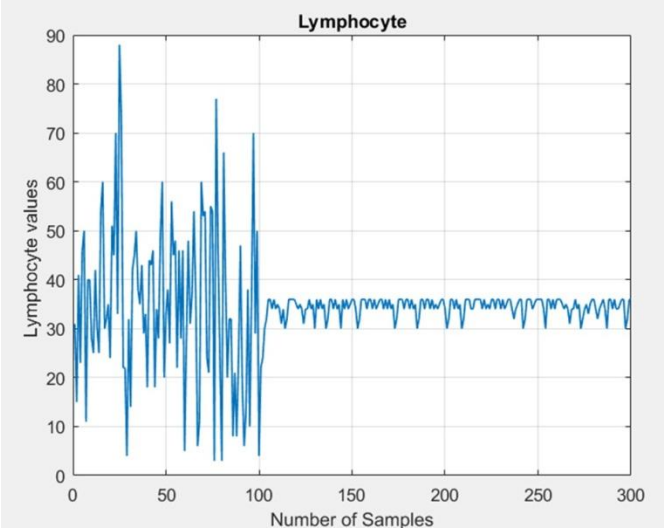


Figure 3.2.j. Lymphocyte count

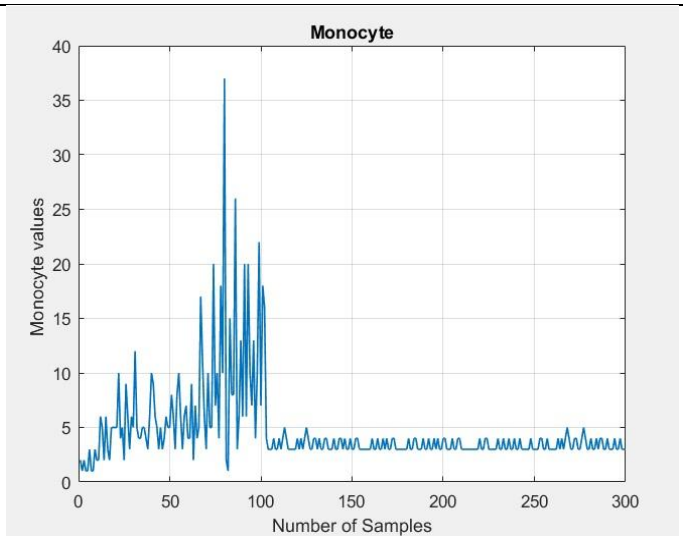


Figure 3.2.k. Monocytes count

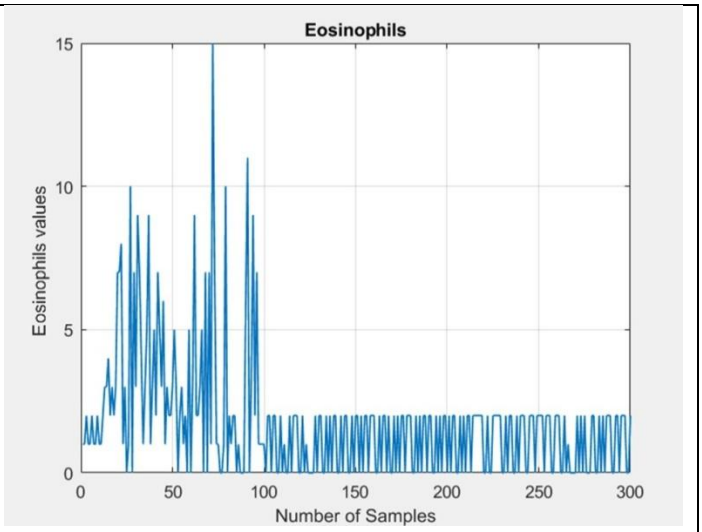


Figure 3.2.l. Eosinophils Count

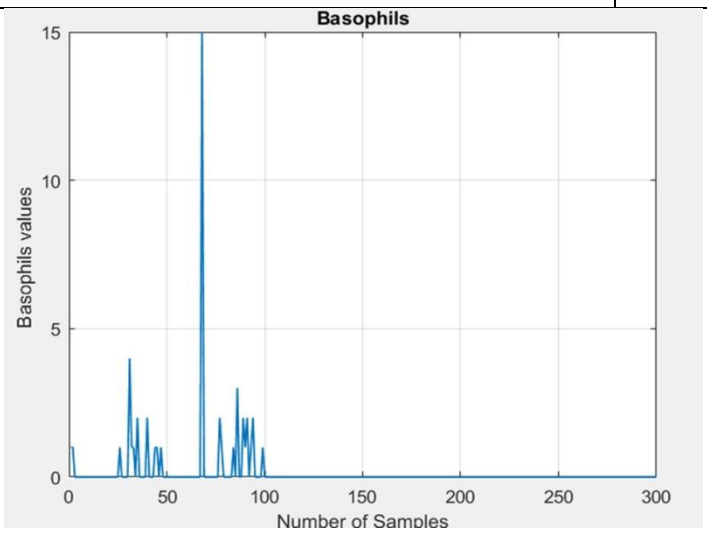


Figure 3.2.m. Basophils count

3.3. Data Pre-processing:

Pre-processing involves ensuring to feed legitimate data for effective clustering process. That plays important role in accuracy of the results. This task is accomplished by handling missing data and addressing any null values in the data set. As we have a data set of 13 parameters of 300 samples, we go by manual inspection to check the presence of irrelevant and null data.

3.4. Conducting K-means clustering:

The model has to be trained. Training involves using the K-means algorithm to learn patterns from the dataset. The model is trained to identify clusters associated with benign and blood cancer conditions. The program is executed in MATLAB IDE for K-means Algorithm. Visualizing clustering results can help understand how data points are grouped or clustered together based on certain similarities or features. Visual scattering output of K-means is shown in the Figures 3.4.a.

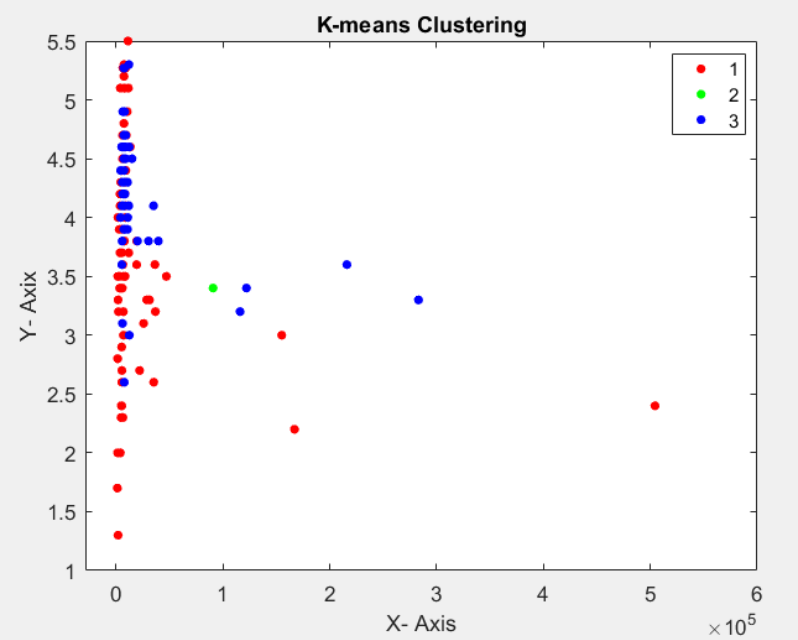


Figure 3.4.a. K-means Clustering

Data is visualized using data points and cluster centroids is shown in Figure 3.4.b.

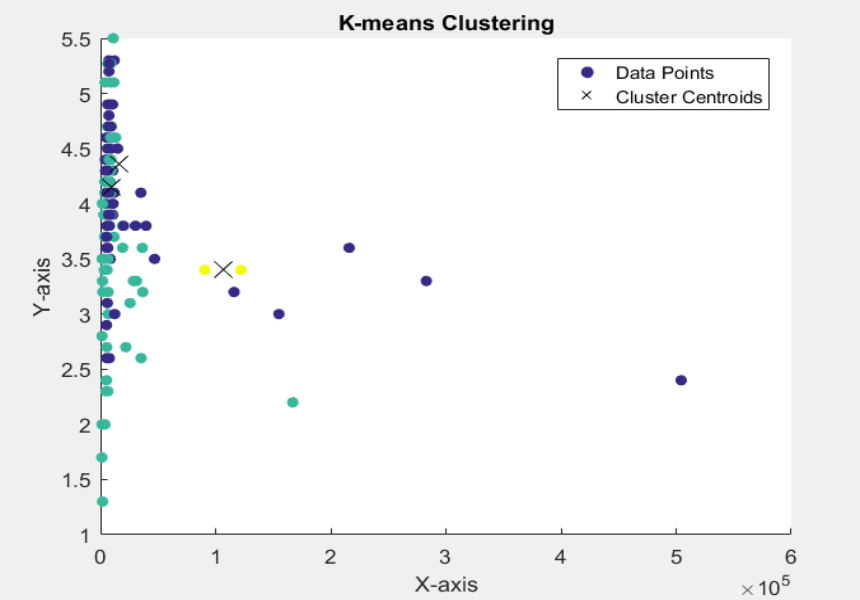


Figure 3.4.b. Data points versus cluster centroids – K means Clustering

In clustering algorithms like K-means, data points and cluster centroids play pivotal roles in the clustering process. Data points refer to individual observations or instances within a dataset, each representing a distinct entity characterized by multiple features or attributes. These points form the dataset and serve as the primary entities on which the clustering algorithm performs its operations. Each data point is typically represented as a vector in the feature space, where each dimension corresponds to a specific attribute or feature of the entity

Cluster Centroids are the representative points within a cluster that summarize or define the characteristics of that cluster. For algorithms like K-means, the centroid of a cluster is the arithmetic mean of all the data points belonging to that cluster. It's a point in the feature space that represents the centre of the cluster. Cluster centroids play a crucial role in algorithms such as K-means, where the goal is to minimize the distances between data points and their assigned centroid. They guide the process of cluster formation by serving as the reference point for grouping data points.

Understanding the nature of data points and cluster centroids is fundamental in comprehending how clustering algorithms group similar data points and determine the representative centres of these groups.

3.5. Prediction:

When a new set of values for any of the thirteen blood parameters is inputted, the trained model will assign these values to the nearest cluster. Based on the cluster assigned, the model will predict whether the patient's condition leans towards benign or indicates potential blood cancer. In case of potential blood cancer, it also predicts in which class it falls.

3.6. Evaluation:

The accuracy and effectiveness of the model in predicting blood cancer are evaluated using metrics like Within-cluster sum of squares (WCSS), Cluster density, Dunn Index and Davies Bouldin Index and Silhouette Score.

IV RESULTS AND DISCUSSIONS

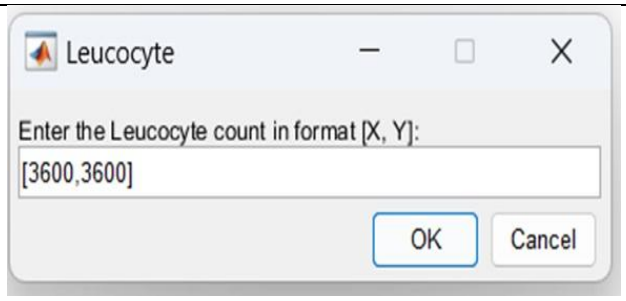
4.1. Findings and Analysis:

We have collected reports from 300 patients, encompassing all five types of diseases. To validate our AI-assisted model, we have specifically chosen reports from five patients, each diagnosed with one of the five types of Leukemia. Below are tables displaying five significant blood parameters:

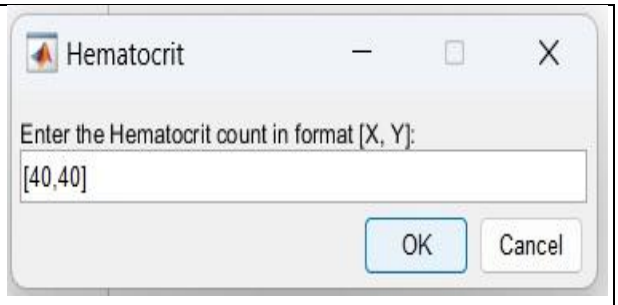
Table 4.1 Parameters of Patient-1

Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
3600	40	31	2	1	Chronic Myeloid Leukemia

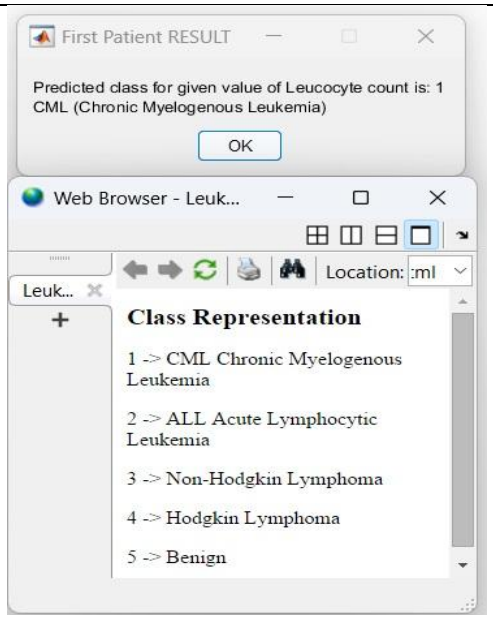
The screenshots of the same while execution during input phase and their output are shown below.



Input value of Leucocyte count (Patien-1)

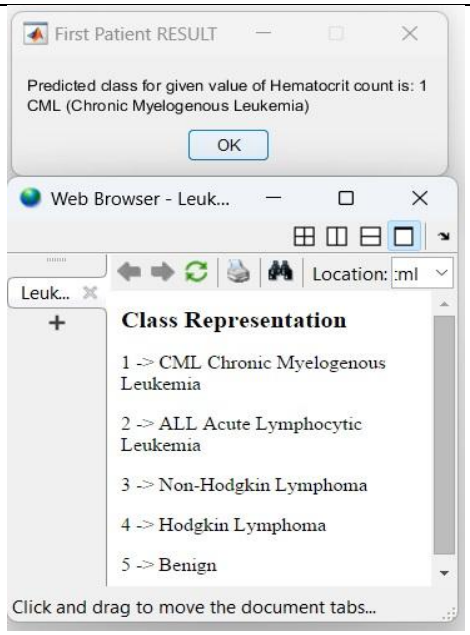


Input value of Hematocrit (Patient-1)



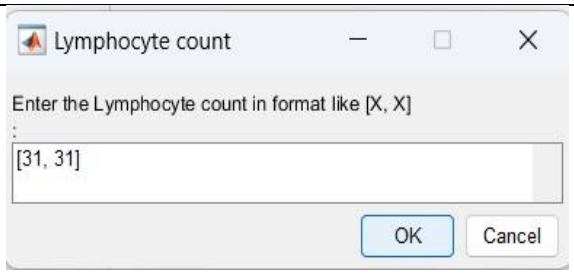
Output predicted class for patient-1

Figure 4.1.a. Input and Output of Leucocyte count (Patient-1)

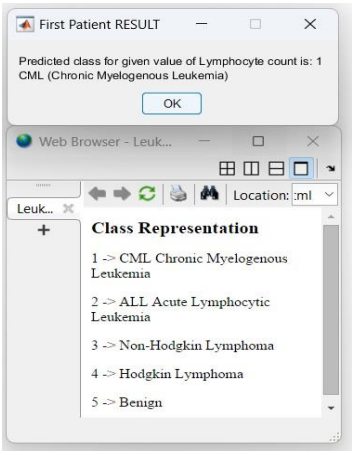


Output predicted class for patient-1

Figure 4.1.b. Input and Output of Hematocrit count (Patient-1)

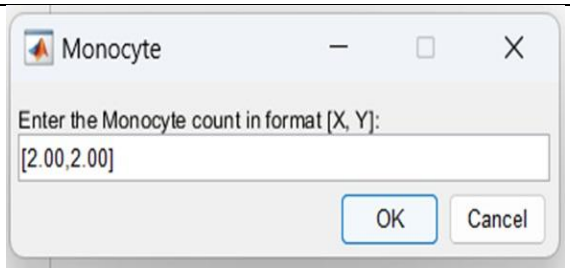


Input value of Lymphocyte Count (Patient-1)

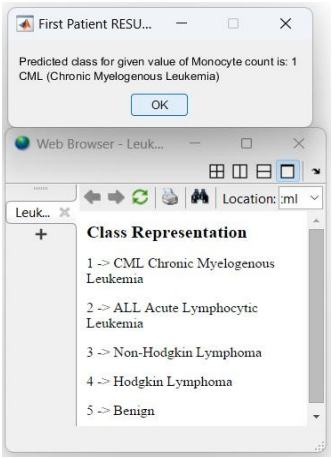


Output predicted class for patient-1

Figure 4.1.c. Input and Output of Lymphocyte count (Patient-1)

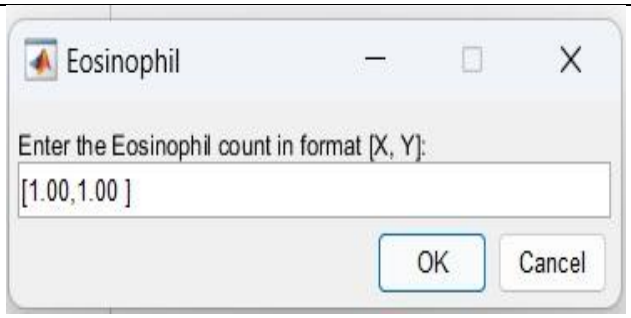


Input value of Monocyte (Patient-1)

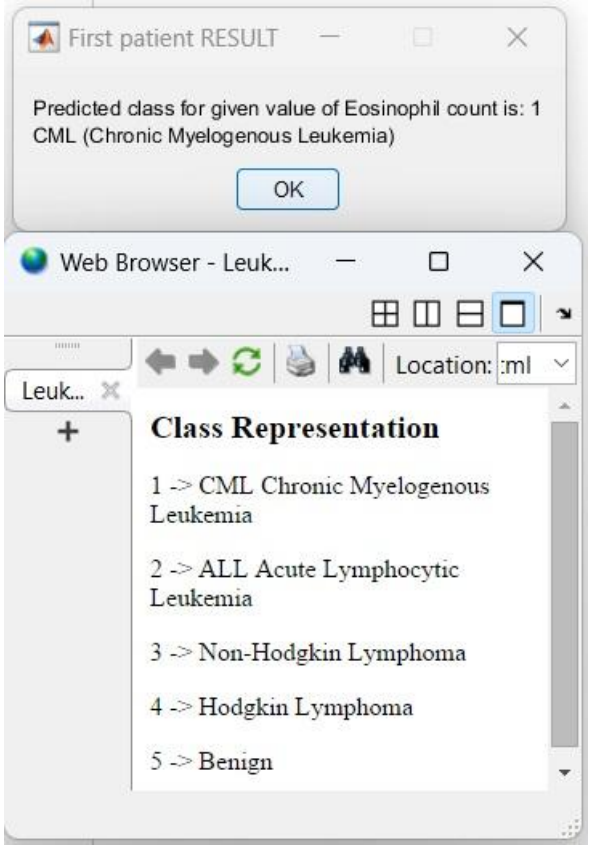


Output predicted class for patient-1

Figure 4.1.d.Input and Output of Monocyte count (Patient-1)



Input value of Eosinophil (Patient-1)

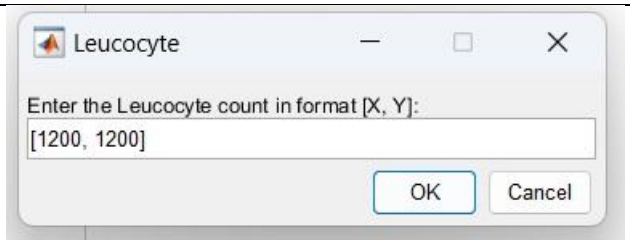


Output predicted class for patient-1

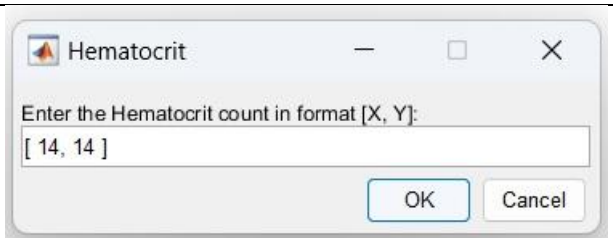
Figure 4.1.e.Input and Output of Eosinophil count (Patient-1)

Table 4.2 Parameters of Patient-2

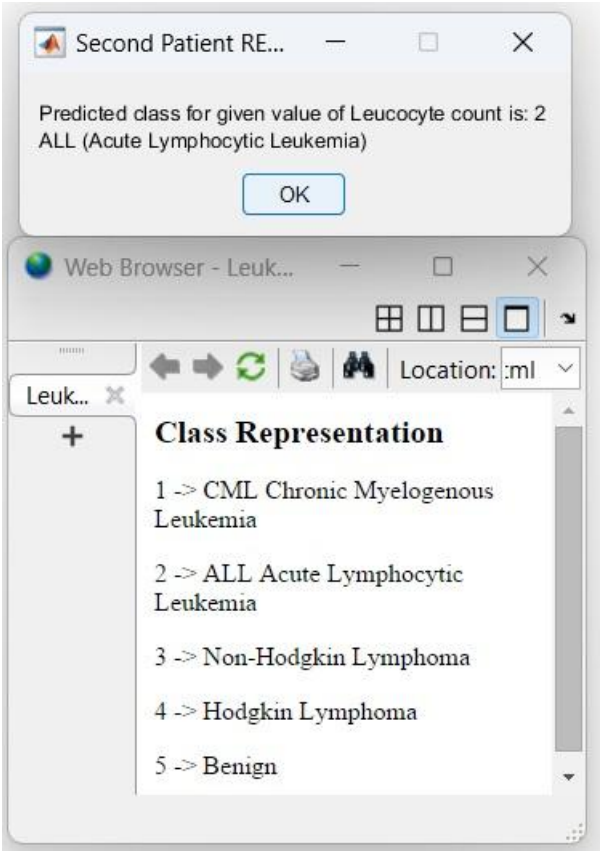
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
1200	14	53	3	7	Acute Lymphocytic Leukemia



Input value of Leucocyte count (Patient-2)

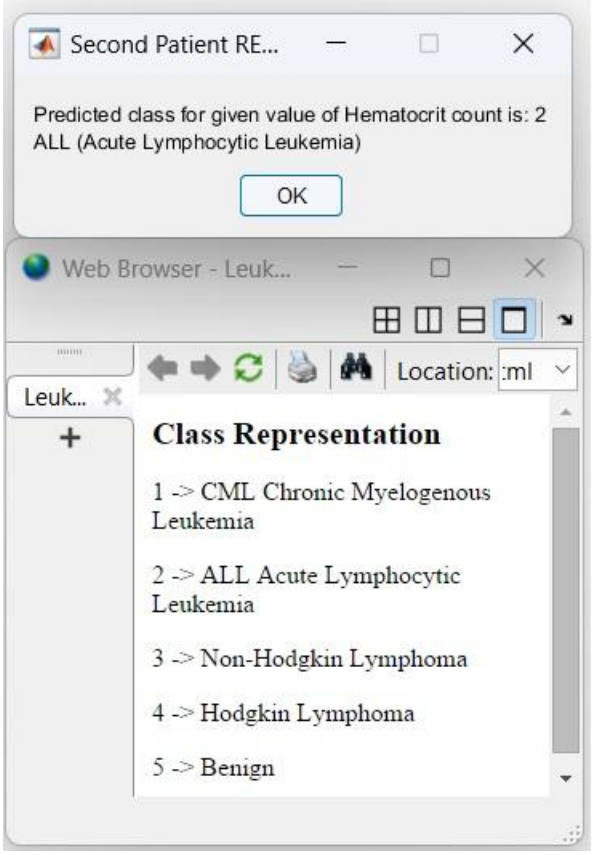


Input value of Hematocrit (Patient-2)



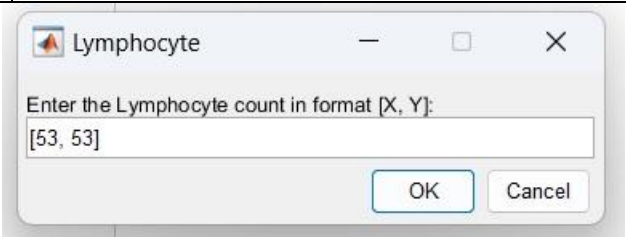
Output predicted class for Patient-2

Figure 4.2.a. Input and Output of Leucocyte count (Patient-2)

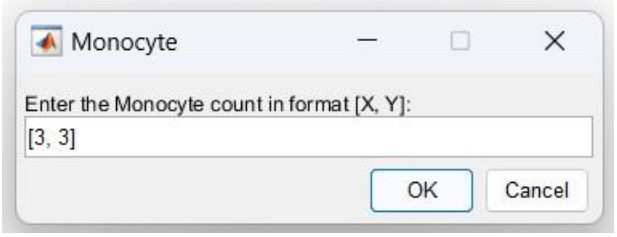


Output predicted class for Patient-2

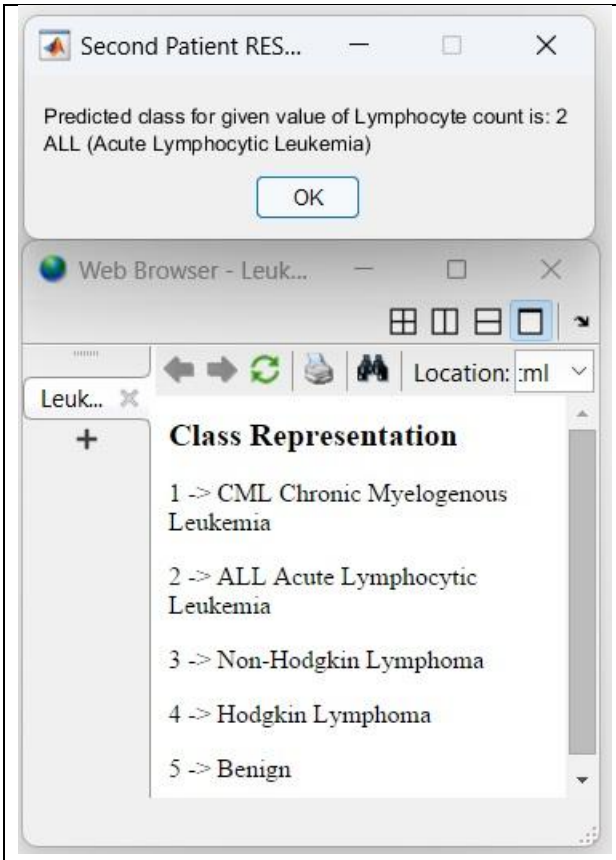
Figure 4.2.b. Input and Output of Hematocrit count (Patient-2)



Input value of Lymphocyte Count (Patient-2)

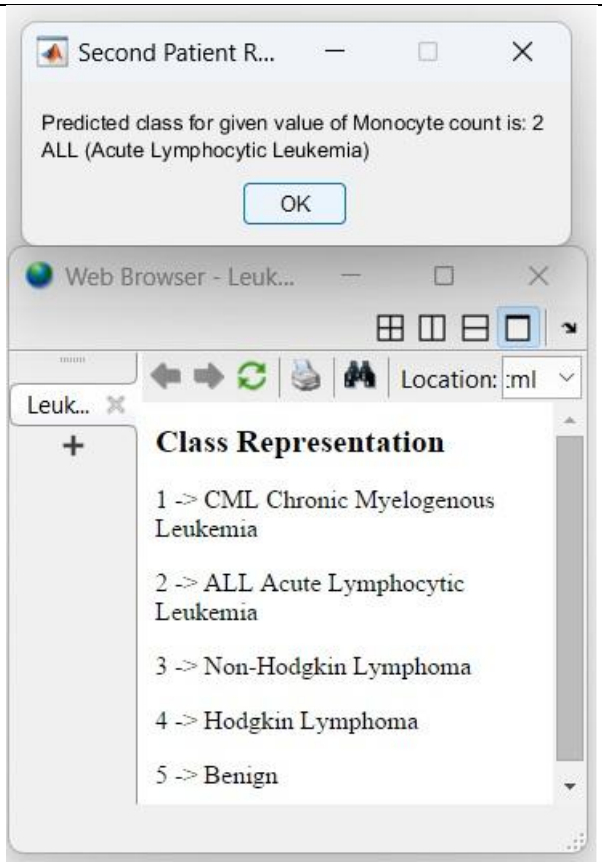


Input value of Monocyte (Patient-2)



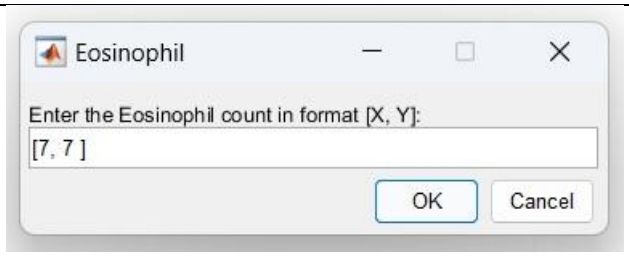
Output predicted class for Patient-2

Figure 4.2.c. Input and Output of Lymphocyte count (Patient-2)

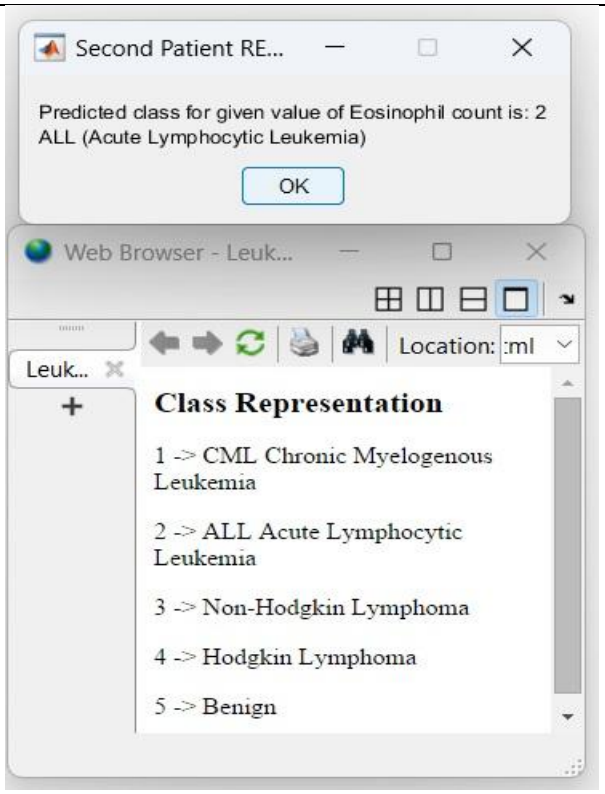


Output predicted class for Patient-2

Figure 4.2.d. Input and Output of Monocyte count (Patient-2)



Input value of Eosinophil (Patient-2)

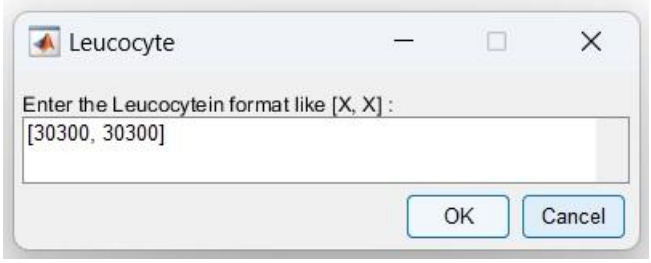


Output predicted class for Patient-2

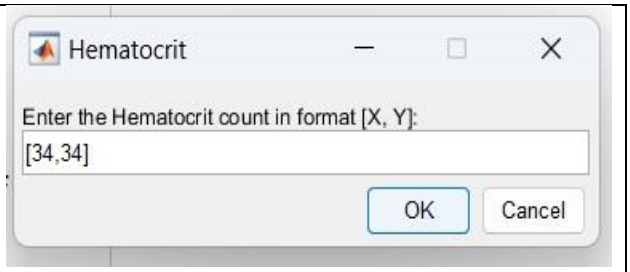
Figure 4.2.e. Input and Output of Eosinophil count (Patient-2)

Table 4.3 Parameters of Patient-3

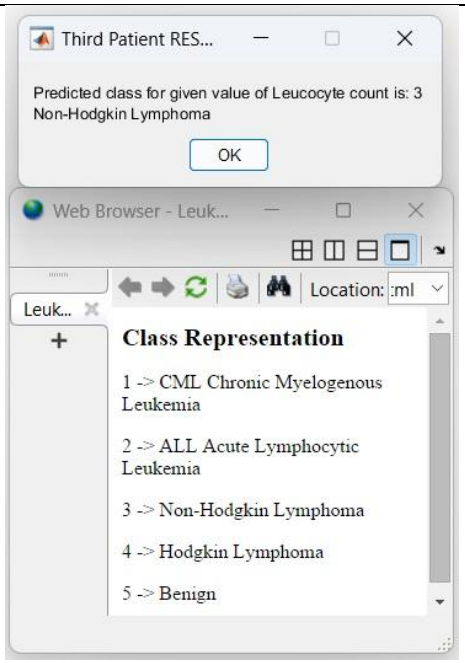
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
30300	34	8	6	11	Non-Hodgkin Lymphoma



Input value of Leucocyte count (Patient-3)

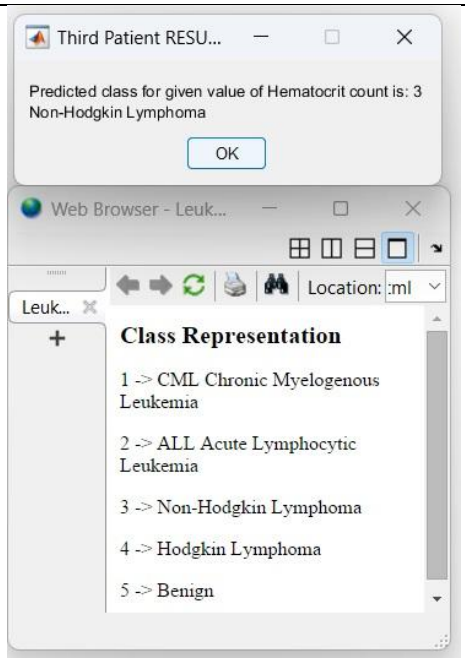


Input value of Hematocrit (Patient-3)



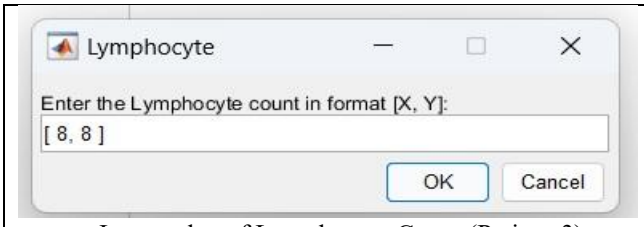
Output predicted class for Patient-3

Figure 4.3.a. Input and Output of Leucocyte count (Patient-3)

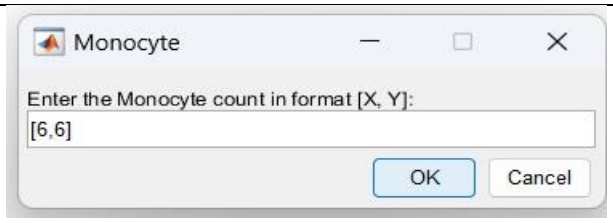


Output predicted class for Patient-3

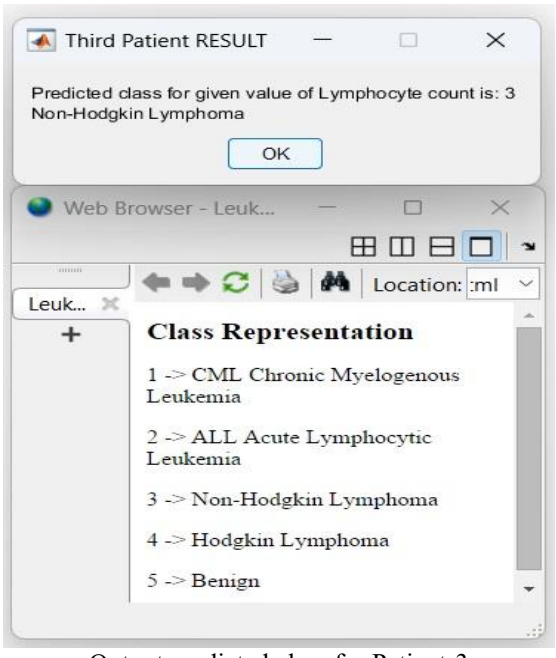
Figure 4.3.b. Input and Output of Hematocrit count (Patient-3)



Input value of Lymphocyte Count (Patient-3)

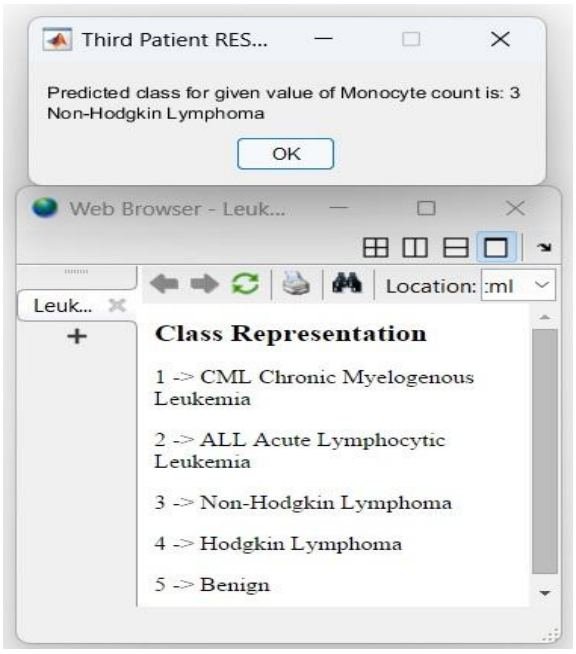


Input value of Monocyte (Patient-3)



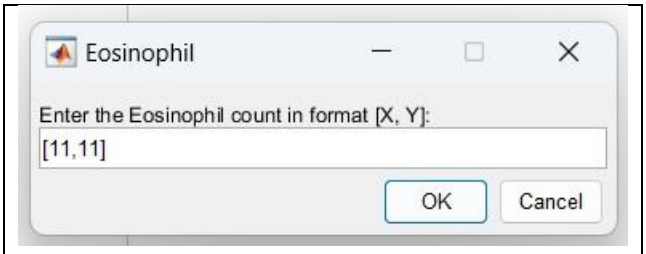
Output predicted class for Patient-3

Figure 4.3.c. Input and Output of Lymphocyte count (Patient-3)

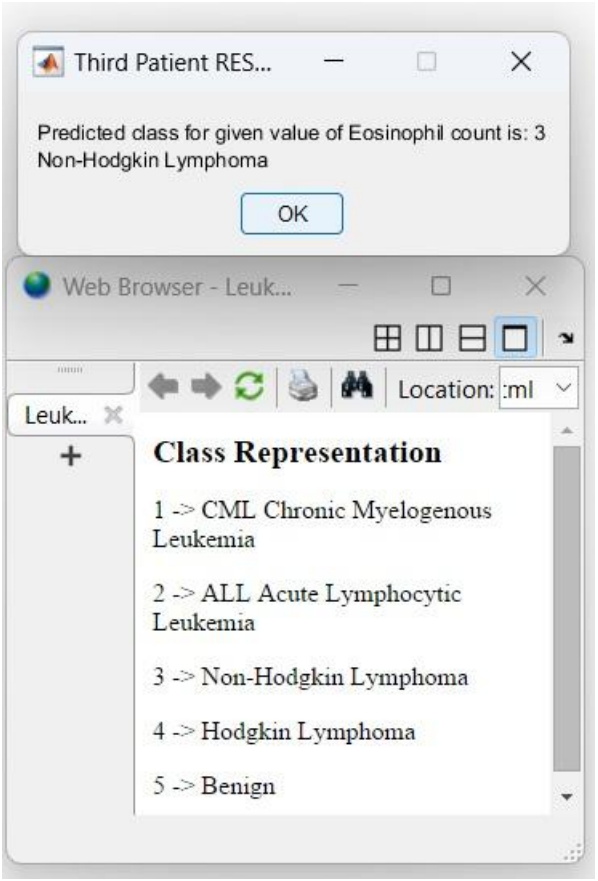


Output predicted class for Patient-3

Figure 4.3.d. Input and Output of Monocyte count (Patient-3)



Input value of Eosinophil (Patient-3)

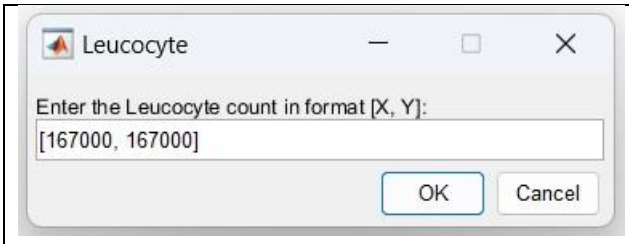


Output predicted class for Patient-3

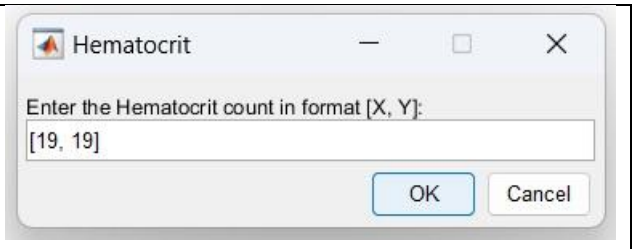
Figure 4.3.e. Input and Output of Eosinophil count (Patient-3)

Table 4.4 Parameters of Patient-4

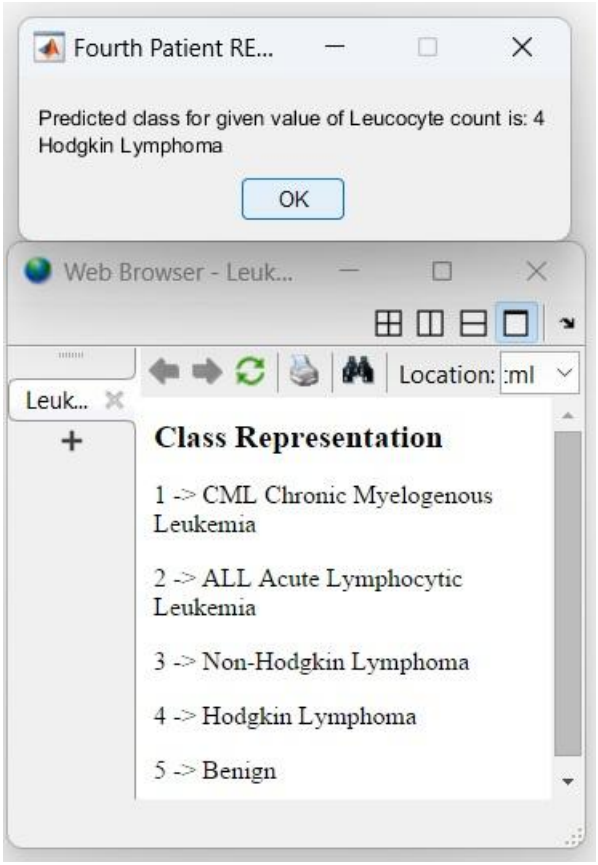
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
167000	19	70	22	1	Hodgkin Lymphoma



Input value of Leucocyte count (Patient-4)

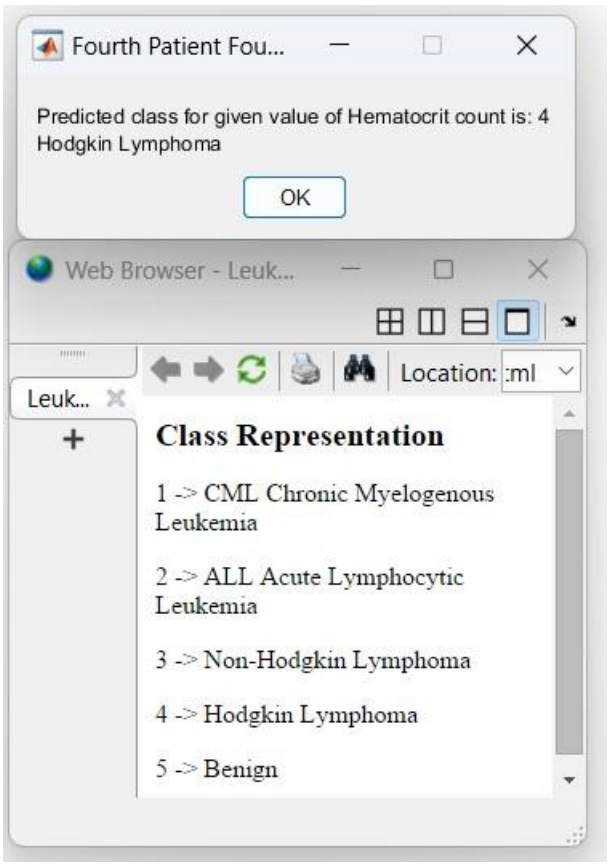


Input value of Hematocrit (Patient-4)



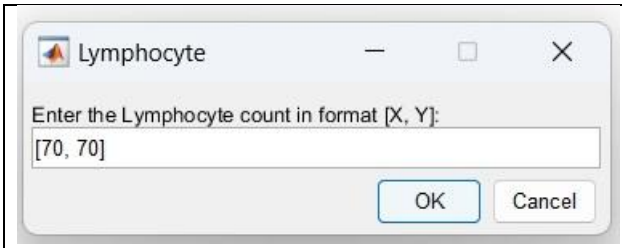
Output predicted class for Patient-4

Figure 4.4.a. Input and Output of Leucocyte count (Patient-4)

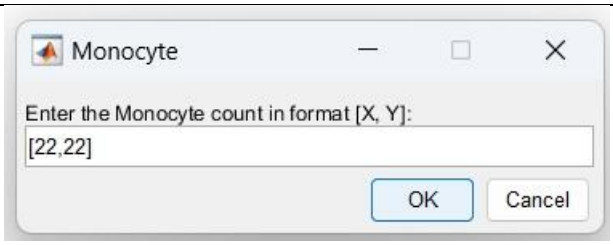


Output predicted class for Patient-4

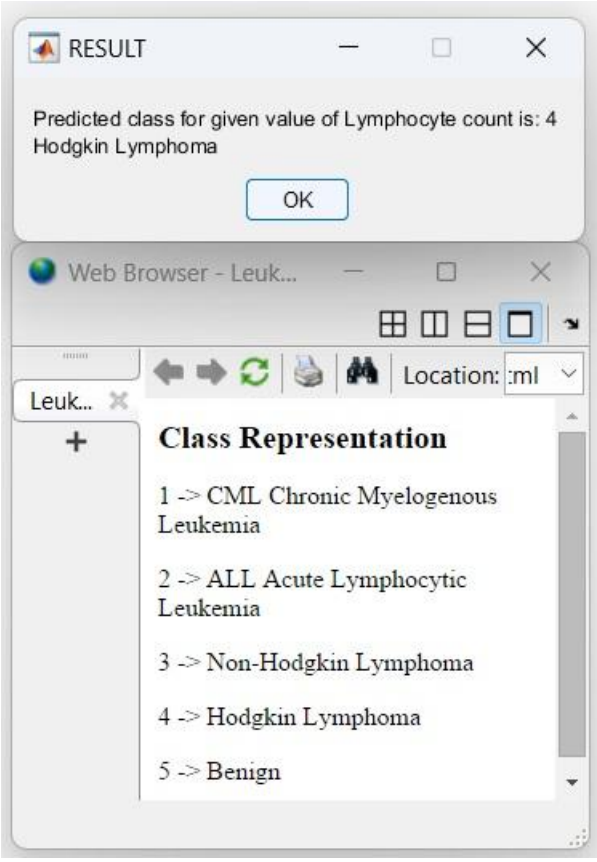
Figure 4.4.b. Input and Output of Hematocrit count (Patient-4)



Input value of Lymphocyte Count (Patient-4)

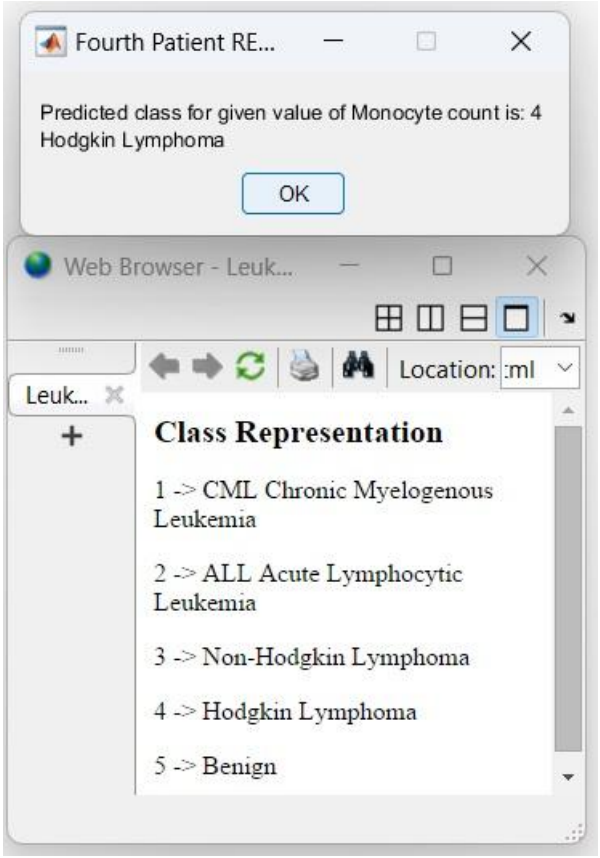


Input value of Monocyte (Patient-4)



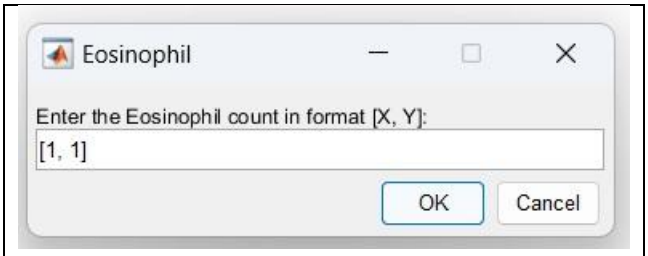
Output predicted class for Patient-4

Figure 4.4.c. Input and Output of Lymphocyte count (Patient-4)

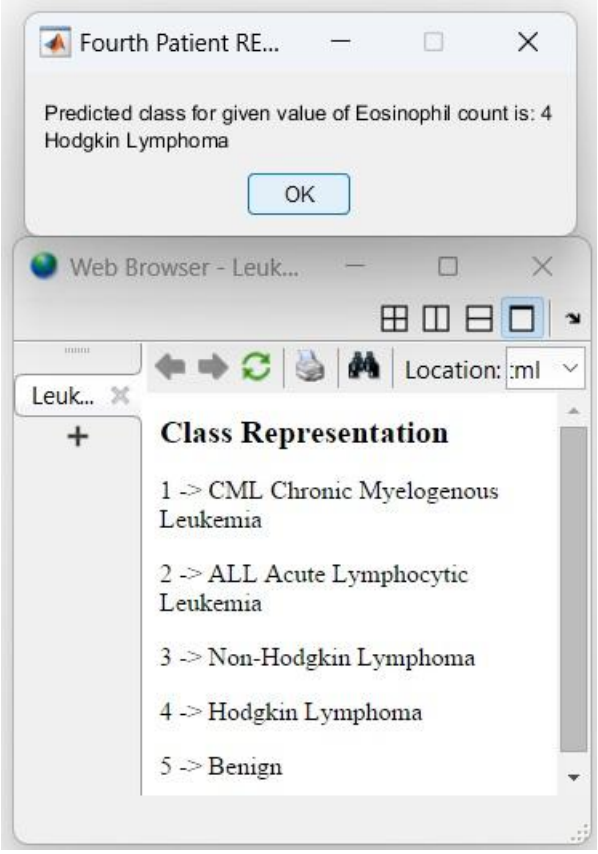


Output predicted class for Patient-4

Figure 4.4.d. Input and Output of Monocyte count (Patient-4)



Input value of Eosinophil (Patient-4)

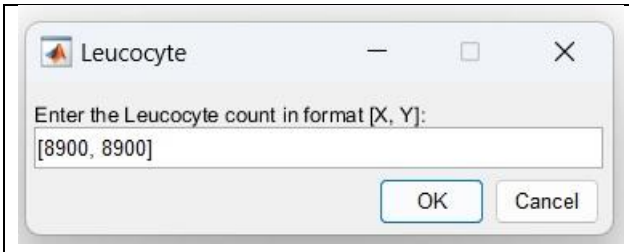


Output predicted class for Patient-4

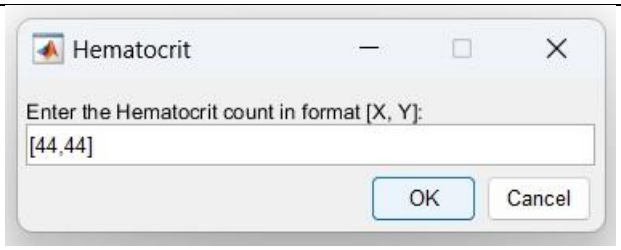
Figure 4.4.e. Input and Output of Eosinophil count (Patient-4)

Table 4.5 Parameters of Patient-5

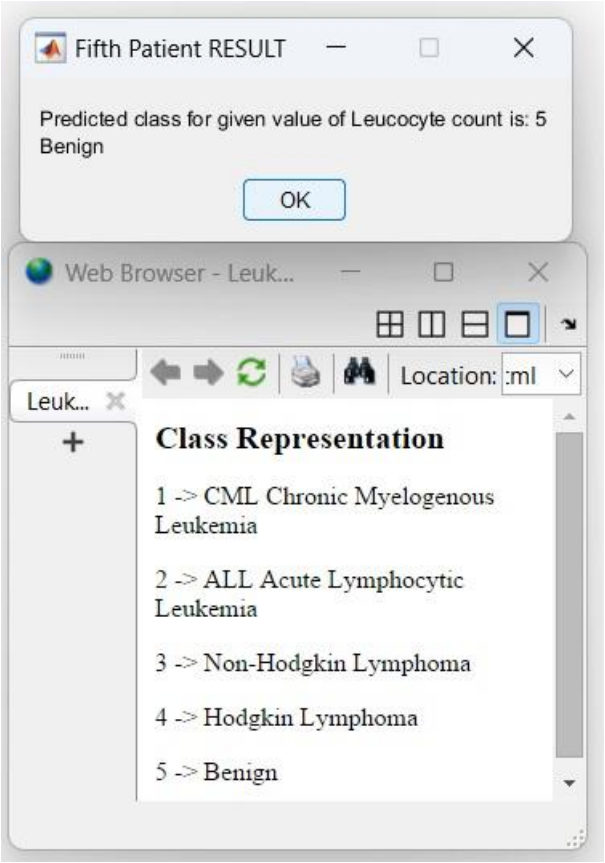
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
8900	44	36	5	2	Benign



Input value of Leucocyte count (Patient-5)

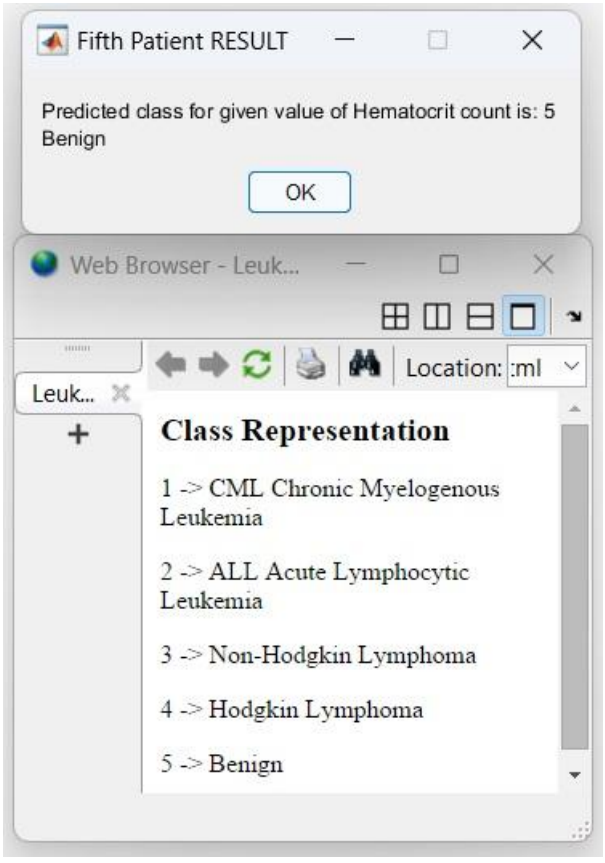


Input value of Hematocrit (Patient-5)



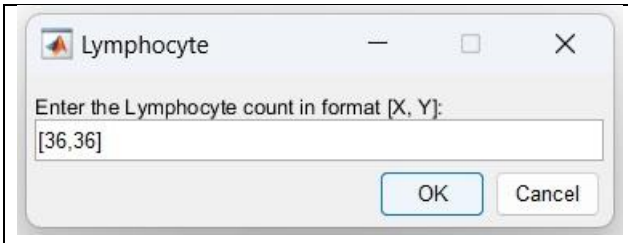
Output predicted class for patient-5

Figure 4.5.a. Input and Output of Leucocyte count (Patient-5)

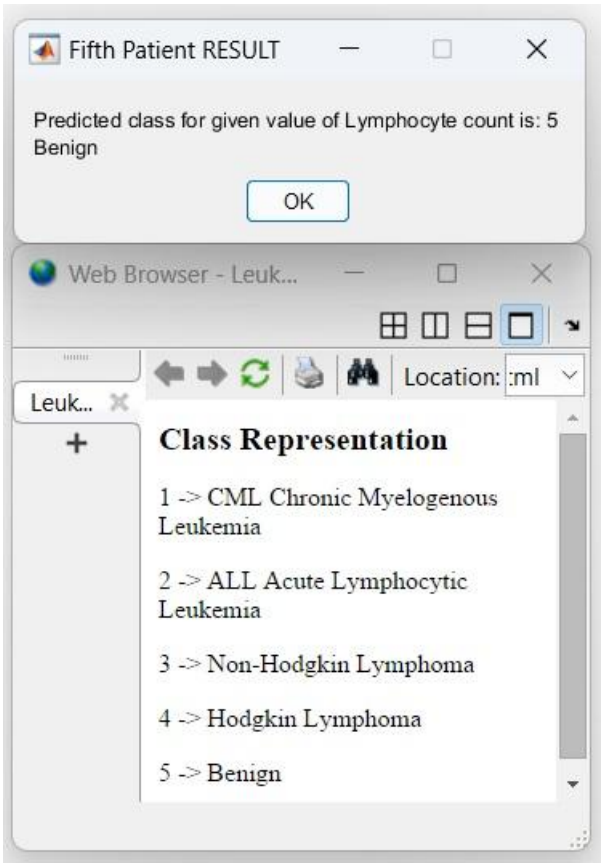


Output predicted class for patient-5

Figure 4.5.b. Input and Output of Hematocrit count (Patient-5)

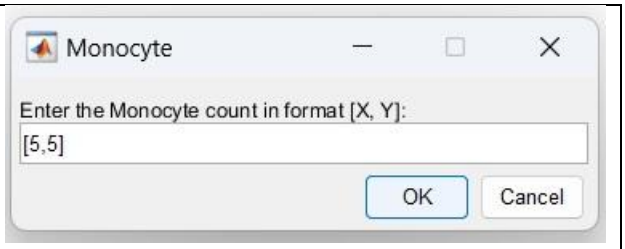


Input value of Lymphocyte Count (Patient-5)

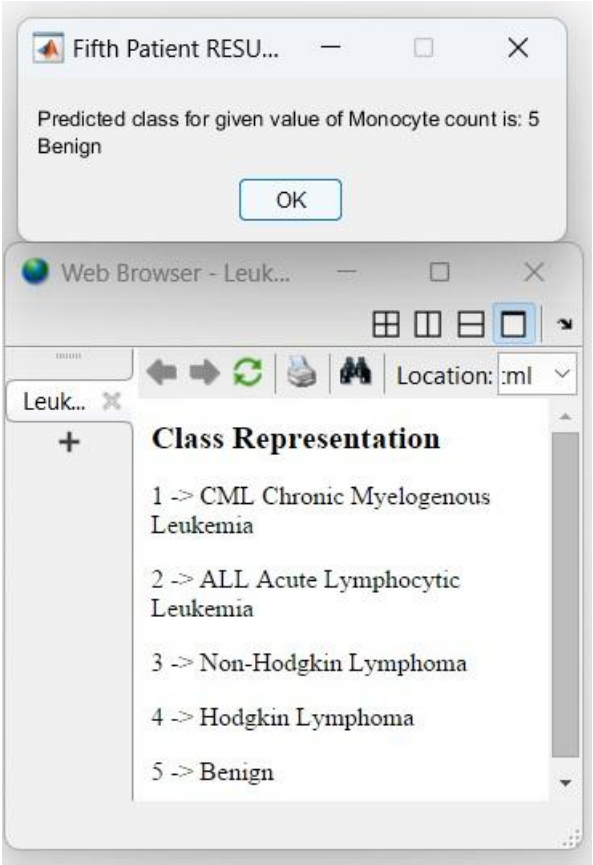


Output predicted class for patient-5

Figure 4.5.c. Input and Output of Lymphocyte count (Patient-5)

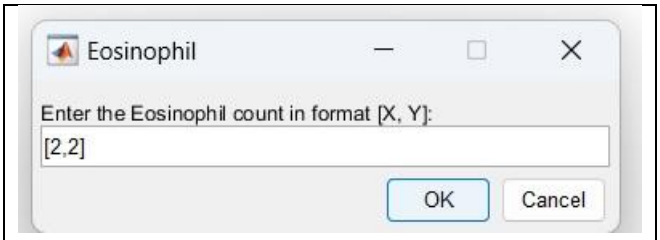


Input value of Monocyte (Patient-5)

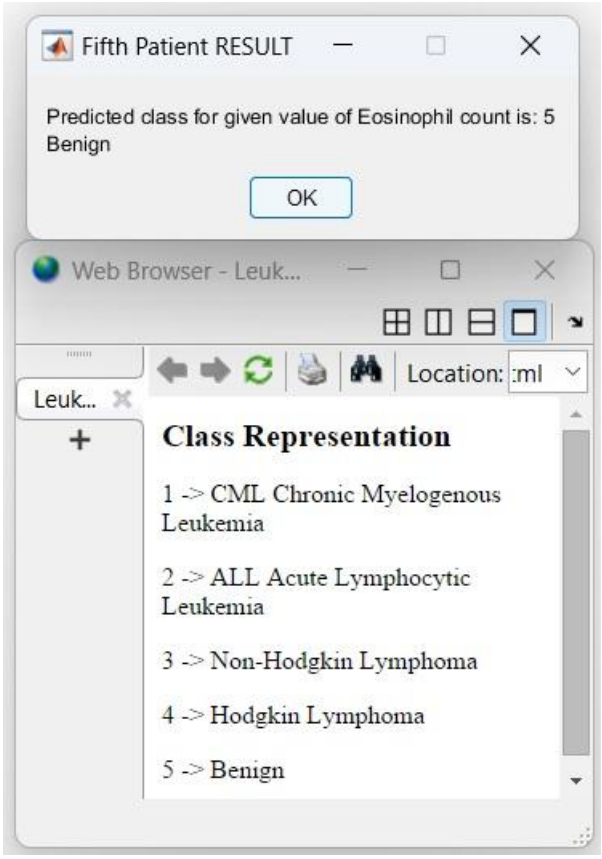


Output predicted class for patient-5

Figure 4.5.d. Input and Output of Monocyte count (Patient-5)



Input value of Eosinophil (Patient-5)



Output predicted class for patient-5

Figure 4.5.e. Input and Output of Eosinophil count (Patient-5)

Thus, the predicted results are quite encouraging as the predicted class falls under the category of the patients infected from the same type of disease.

4.2 Performance evaluation:

In order to confirm the reliability of algorithm, K-means clustering algorithm is evaluated using most basic metrics such as accuracy, precision and recall. Accuracy measures the proportion of correctly classified data points compared to the total number of data points. In the context of K-means clustering, accuracy can be calculated by comparing the cluster assignments produced by the algorithm to the true labels of the data[45]

The Accuracy versus number of neighbours(k) graph is shown in the figure 4.2.a. below. The value of accuracy for the values of neighbours 3-5 k is around 85% which is quite satisfying.

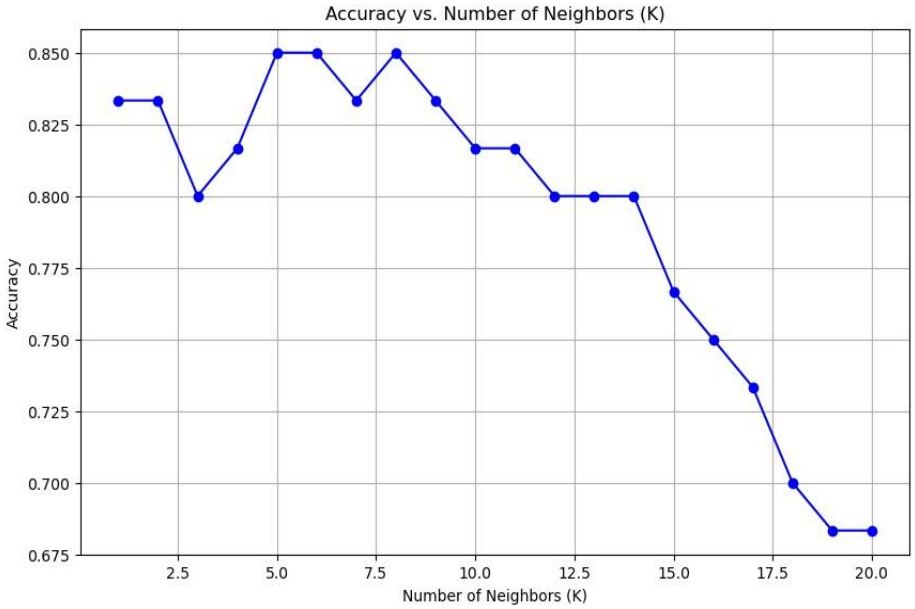


Figure 4.2.a. Accuracy versus Number of Neighbour (k)

Precision and recall are widely used evaluation metrics in the realm of classification algorithms to gauge their effectiveness. Precision specifically measures the accuracy of positive predictions made by a classifier. It quantifies the ratio of correctly predicted positive instances to the total instances predicted as positive. In essence, precision assesses the model's capability to avoid mislabeling negative samples as positive [45]

Precision = True Positives / (True Positives + False Positives)

Recall, also known as sensitivity or true positive rate, measures the ratio of correctly predicted positive observations to the actual positives in the dataset. It quantifies the ability of the model to identify all relevant instances (true positives) within a dataset.

Recall = True Positives / (True Positives + False Negatives)

As shown in the Figure 4.2.b the precision is about 83% and recall is 85%, which is quite satisfying.

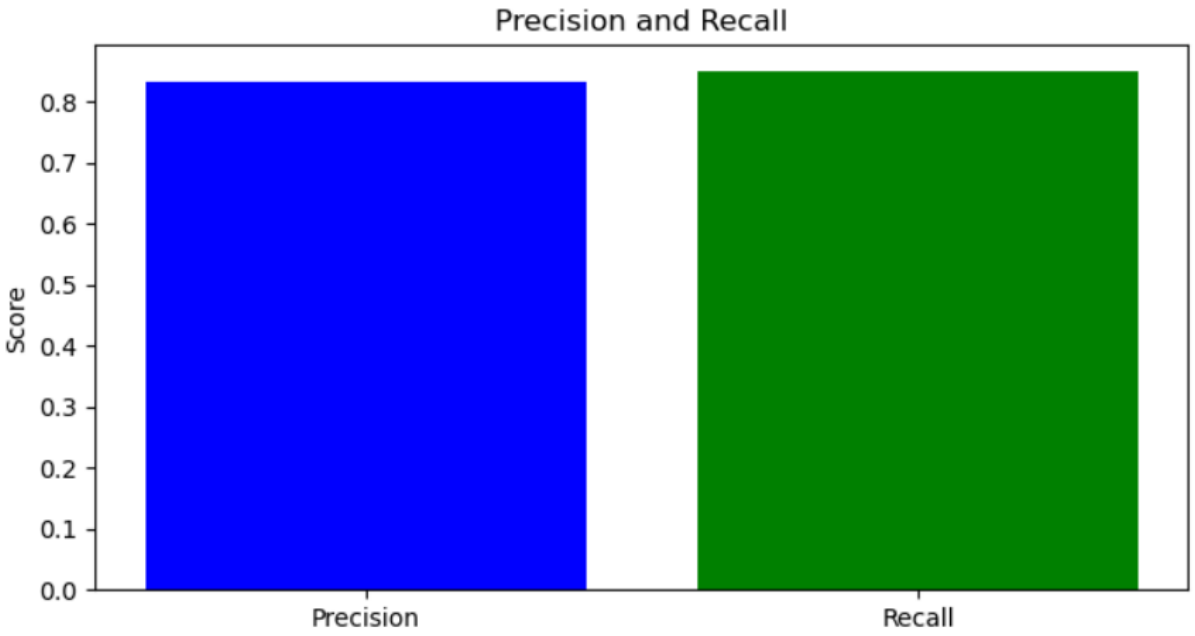


Figure 4.2.b. Precision and Recall

V CONCLUSION

Leukemia poses a significant threat as a blood cancer disease, often leading to fatal outcomes. Timely detection is crucial for effective treatment and patient care. Automated systems, particularly those employing machine learning algorithms, offer promising avenues for supporting pathologists in the blood diagnosis. The integration of AI technologies holds great potential for advancing medical diagnostics. In this study, an AI-assisted program successfully predicted five classes of cancer, based on blood parameters, with the K-means algorithm demonstrating effective performance as evaluated through metrics including accuracy, precision, and recall.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author, [Tabasum Guledgudd]. The data are not publicly available due to restrictions that could compromise the privacy concerns.

DISCLOSURE STATEMENT

The authors have no relevant financial or non-financial competing interests to report.

FUNDING

The study was carried out without receiving financial assistance from external sources

References

[1] Blackadar CB. Historical review of the causes of cancer. World J Clin Oncol. 2016 Feb 10;7(1):54-86. DOI: 10.5306/wjco.v7.i1.54. PMID: 26862491; PMCID: PMC4734938.

[2] Tabasum Guledgudd, Noorullah Shariff, S.A. Quadri and SayedAbdulhayan, Leukemia Disease: Overview and Detection Approaches, Journal of Systems Engineering and Electronics (ISSN NO: 1671-1793) Vol: 34 Issue 6, pp.126-142, 2024, https://jseepublisher.com/wpcontent/uploads/14-JSEE2333.pdf , (SCOPUS).

[3] Tabasum Guledgudd, Noorullah Shariff and S.A. Quadri, A Comparative Study of K-Means, GMM, SVM, and Random Forest for Enhancing Machine Learning in Leukemia Diagnosis, African Journal of biomedical Research, Vol. 27(4s) (November 2024); 4257-4268, (ISSN: 1119-5096, SCOPUS - Q3), DOI: https://doi.org/10.53555/AJBR.v27i4S.4392.

[4] Tabasum Guledgudd, Noorullah Shariff C & Sayed Abulhasan Quadri "A comprehensive review: State of art integrated technologies in IoHT applications" DOI:10.1080/20421338.2024.2417447. (SCOPUS).

[5] Cheon, Hwayeong, Jin Ho Paik, Moran Choi, Hee-Jin Yang, and Joo-Hiuk Son. "Detection and manipulation of methylation in blood cancer DNA using terahertz radiation." *Scientific reports* 9, no. 1 (2019): 6413.

[6] Xing, Wenchao, and Yilin Bei. "Medical health big data classification based on KNN classification algorithm." *IEEE Access* 8 (2019): 28808-28819. https://doi.org/10.1109/ACCESS.2019.2955754

[7] Safuan, Syadia Nabilah Mohd, Tomari, Wan Nurshazwani Wan Zakaria, Mohd NorzaliHj Mohd, and Nor Surayahani Suriani. "Investigation of white blood cell biomaker model for acute lymphoblastic leukemia detection based on convolutional neural network." *Bulletin of Electrical Engineering and Informatics* 9, no. 2 (2020): 611-618. https://doi.org/10.11591/eei.v9i2.1857

[8] Nighat, Misba Sikandar, Ikram Ud Din, Ahmad Almogren, and Sikandar Ali. "IoMT-based automated detection and classification of leukemia using deep learning." *Journal of healthcare engineering* 2020 (2020): 1-12. https://doi.org/10.1155/2020/6648574

[9] Jabeen Aqsa,Sara Jabeen; Syed Adnan Shah, "Efficient Features for Effectively Detection of Leukemia Cells", Wakeel Ahmad Rao, 2020 IEEE 23rd International Multitopic Conference (INMIC), Year: 2020, Conference Paper, Publisher: IEEEhttps://doi.org/10.1109/INMIC50486.2020.9318085

[10] Hossain Mohammad Akter, Mubtasim Rahman, Shadikun Nahar Sakiba,. Muzahidul Islam,, Ashir Ahmed, “An effective leukemia prediction technique using supervised machine learning classification algorithm”, Advanced Information and Communication Technology, Proceedings of International Conference on Trends in Computational and Cognitive Engineering - Proceedings of TCCE 2020

[11] Genovese, Angelo, Mahdi S. Hosseini, Vincenzo Piuri, Konstantinos N. Plataniotis, and Fabio Scotti. "Acute lymphoblastic leukemia detection based on adaptive unsharpening and deep learning." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1205-1209. IEEE, 2021. https://doi.org/10.1109/ICASSP39728.2021.9414362 Aftab, Muhammad Omer, Mazhar Javed Awan, Shahid Khalid, Rabia and Hassan

[12] Shabir. "Executing spark BigDL for leukemia detection from microscopic images using transfer learning." In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pp. 216-220. IEEE, 2021. https://doi.org/10.1109/CAIDA51941.2021.9425264.

[13] Pałczyński, Krzysztof, Sandra Śmigiel, Marta Gackowska, Damian Ledziński, Sławomir Bujnowski, and Zbigniew Lutowski. "IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification." *Sensors* 21, no. 23 (2021): 8025. https://doi.org/10.3390/s21238025

[14] Desi, A., Kumar, R., and Sharma, P., “Automated optical image processing system for leukemia detection and white blood cell classification,” *Journal of Medical Imaging and Health Informatics*, vol. 11, no. 3, pp. 745–752, 2021.

[15] Hossain, Mohammad Akter, Mubtasim Islam Sabik, Md Moshiur Rahman, Shadikun Nahar Sakiba, A. K. M. Muzahidul Islam, SwakkharShatabda, Salekul Islam, and Ashir Ahmed. "An effective leukemia prediction technique using supervised machine learning classification algorithm." In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*, pp. 219-229. Springer Singapore, 2021.

[16] Armya, Revella EA, Adnan Mohsin Abdulazeez, Amira Bibo Sallow, and Diyar Qader Zeebaree. "Leukemia Diagnosis using Machine Learning Classifiers Based on Correlation Attribute Eval Feature Selection." Asian Journal of Research in Computer Science (2021): 52-65. DOI: 10.9734/AJRCOS/2021/v9i330225

[17] Diana Passaro, Manuel, Giovanni, Probir, Linda Ariza- McNaughton, Clara Borràs-Eroles "Integrated OMICs unveil the bone-marrow microenvironment in human leukemia." Cell reports 35, no. 6 (2021). https://doi.org/10.1016/j.celrep.2021.109119

[18] Jusman, Yessi, Slamet Riyadi, Amir Faisal, Siti Nurul Aqmariah Mohd Kanafiah, Zeehaida Mohamed, and Rosline Hassan. "Classification system for leukemia cell images based on hu moment invariants and support vector machines." In 2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 137-141. IEEE, 2021. https://doi.org/10.1109/ICCSCE52189.2021.9530974

[19] Kumar, T. Ananth, K. Suresh Kumar, and A. Jayakumar. "Analysis of IoT based Leukemia Detection Techniques." (2022).

[20] Zhong, Pengqiang, Mengzhi Hong, He, Zhang, Yaoming Chen, Wang, Peisong Chen, and Juan Ouyang. "Diagnosis of Acute Leukemia by Multiparameter Flow Cytometry with the Assistance of Artificial Intelligence." *Diagnostics* 12, no. 4 (2022): 827. https://doi.org/10.3390/diagnostics12040827

[21] Hossain, Mohammad Akter, AKM Muzahidul Islam, Salekul Islam, SwakkharShatabda, and Ashir Ahmed. "Symptom based explainable artificial intelligence model for leukemia detection." *IEEE Access* 10 (2022): 57283-57298. https://doi.org/10.1109/ACCESS.2022.3176274

[22] Ananth, Christo, K. Mystica, M. Sridharan, S. Aaron James, and T. Ananth Kumar. "An Advanced Low-cost Blood Cancer Detection System." *International Journal of Early Childhood Special Education* 14, no. 5 (2022). DOI:10.9756/INTJECSE/V14I5.475

[23] Sridhar, K., Ajay Reddy Yeruva, P. N. Renjith, Asmita Dixit, Aatif Jamshed, and Ravi Rastogi. "Enhanced Machine learning algorithms Lightweight Ensemble Classification of Normal versus Leukemic Cel." *Journal of Pharmaceutical Negative Results* (2022): 496-505. https://doi.org/10.47750/pnr.2022.13.S09.056 Das, Pradeep Kumar, V. A. Diya, Sukadev Meher, Panda, and Ajith Abraham. "A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia." *IEEE access* (2022). https://doi.org/10.1109/ACCESS.2022.3196037

[24] Hossain, Mohammad Akter, AKM Muzahidul Islam, Salekul Islam, SwakkharShatabda, and Ashir Ahmed. "Symptom based explainable artificial intelligence model for leukemia detection." *IEEE Access* 10 (2022): 57283-57298.

[25] Devi, Tulasi Gayatri, Nagamma Patil, Sharada Rai, and Cheryl Sarah Philipose. "Gaussian blurring technique for detecting and classifying acute lymphoblastic leukemia cancer cells from microscopic biopsy images." *Life* 13, no. 2 (2023): 348. https://doi.org/10.3390/life13020348

[26] Khademi, Reyhane, Zahra Mohammadi, Rahele Khademi, Amene Saghazadeh, and Nima Rezaei. "Nanotechnology-based diagnostics and therapeutics in acute lymphoblastic leukemia: a systematic review of preclinical studies." *Nanoscale Advances* (2023). https://doi.org/10.1039/D2NA00483F Della Starza, Irene, Lucia Anna, Loredana Elia, Vittorio Bellomarino, Marco

[27] Beldinanzi, Roberta Soscia, Deborah Cardinali, Sabina Chiaretti, Anna Guarini, and Robin Foà. "Optimizing molecular minimal residual disease analysis in adult acute lymphoblastic leukemia." *Cancers* 15, no. 2 (2023): 374. https://doi.org/10.3390/cancers15020374 Varadarajan, Indumathy, Eric Pierce, Scheuing, Amy Morris, Firas El Chaer, and

[28] Michael Keng. "Post-Hematopoietic Cell Transplantation Relapsed Acute Lymphoblastic Leukemia: Current Challenges and Future Directions." *OncoTargets and Therapy* (2023): 1-16.

[29] Hoffmann (b), Joerg, Michael C. Thrun, Maximilian A. Röhnert, Malte von Bonin, Uta Oelschlägel, Andreas Neubauer, Alfred Ultsch, and Cornelia Brendel. "Identification of critical hemodilution by artificial intelligence in bone marrow assessed for minimal residual disease analysis in acute myeloid leukemia: The Cinderella method." *Cytometry Part A* 103, no. 4 (2023): 304-312. https://doi.org/10.1002/cyto.a.24686

[30] Liu, Hui. "SETD2 detection may reveal response to induction therapy and survival profile in acute myeloid leukemia patients." *Hematology* 28, no. 1 (2023): 2161194. https://doi.org/10.1080/16078454.2022.2161194

[31] Tiago, Brito-Rocha, Vera Constâncio, Rui Henrique, and Carmen Jerónimo. "Shifting the cancer screening paradigm: the rising potential of blood-based multi-cancer early detection tests." *Cells* 12, no. 6 (2023): 935. https://doi.org/10.3390/cells12060935

[32] PierPiccaluga, Paolo, Stefania Paolini, and Giuseppe Visani. "Antigen Receptors Gene Analysis for Minimal Residual Disease Detection in Acute Lymphoblastic Leukemia: The Role of High Throughput Sequencing." *Hemato* 4, no. 1 (2023): 42-55. https://doi.org/10.3390/hemato4010004

[33] Naveed, Muhammad, Noor ul Ain, Tariq Aziz, Khushbakht Javed, Muhammad Aqib Shabbir, Metab Alharbi, Abdulrahman Alsahammari, and Abdullah F. Alasmari. "Artificial Intelligence Assisted Pharmacophore Design for Philadelphia Chromosome-Positive Leukemia with Gamma-Tocotrienol: A Toxicity Comparison Approach with Asciminib." *Biomedicines* 11, no. 4 (2023): 1041. https://doi.org/10.3390/biomedicines11041041

[34] Molin,Dal, Anna, Caterina TrettiParenzan, Enrico Gaffo, Cristina Borin, Elena Boldrin, Lueder H. Meyer, Geertruij TeKronnie, Silvia Bresolin, and Stefania Bortoluzzi. "Discovery of fusion circular RNAs in leukemia with KMT2A::AFF1 rearrangements by the new software CircFusion." *Briefings in Bioinformatics* 24, no. 1 (2023): bbac589. https://doi.org/10.1093/bib/bbac589

[35] Ulya,Farahdina, Amastasia Salsabila Muliawati, Zulfa, MiftakhulFirdhaus, Ihwanul Aziz, Hari Suprihatin, Darsono Darsono, Nasori Nasori, and Agus Rubiyanto. "Electrochemical and Optical Analysis of Various Compositions of Au and Ag Layers for Blood Cancer Prognosis." *Coatings* 13, no. 1 (2023): 186. https://doi.org/10.3390/coatings13010186

[36] Alejandro,Hortal Marta Lacuna, Claudia Cifuentes, Miguel Alcoceba, Xosé R. Bustelo, Marcos González, and Balbino Alarcón. "An Optimized Single Nucleotide Polymorphism-Based Detection Method Suggests That Allelic Variants in the Untranslated Region of RRAS2 Correlate with Treatment Response in Chronic Lymphocytic Leukemia Patients." *Cancers* 15, no. 3 (2023): 644. https://doi.org/10.3390/cancers15030644

[37] Ahmad, W.; Iqbal, M.; Amin, M.A.; Bangyal, W.H.; Shahzad, A.R. MachineLearning Driven Dashboard for Chronic Myeloid Leukemia Prediction using Protein Sequences. Preprints 2023, 2023120053. https://doi.org/10.20944/preprints202312.0053.v1.

[38] Nasir, Muhammad Umar, Muhammad Farhan Khan, Muhammad Adnan Khan, Muhammad Zubair, Sagheer Abbas, Meshal Alharbi, and Md Akhtaruzzaman. "Hematologic Cancer Detection Using White Blood Cancerous Cells Empowered with Transfer Learning and Image Processing." Journal of Healthcare Engineering 2023. https://doi.org/10.1155/2023/1406545 Jules, Chassetuillier,Meriem Benmerad, Pierre-Louis Lecoq, Romain Favier, Matthieu

[39] Trancart, Sandrine Hayette, Vincent Alcazer et al. "Identification of Individual Risk Factors for Treatment Free Remission (TFR) Failure in Chronic Phase Chronic Myeloid Leukemia (CP-CML) Patients: Predictive Models Developed on a Nationwide Clinical and Claim Database." Blood 142 (2023): 6369. https://doi.org/10.1182/blood-2023-179970 Elhadary, Mohamed, Ahmed Adel Elsabagh, Khaled Ferih, Basel Elsayed, Amgad M.

[40] Elshoeibi, Rasha Kaddoura, Akiki, Khalid Ahmed, and Mohamed Yassin. "Applications of machine learning in chronic myeloid leukemia." Diagnostics 13, no. 7 (2023): 1330. https://doi.org/10.3390/diagnostics13071330

[41] Penmetsa, Venkat, Lachin Fernando, and Sindhu Ghanta. "Improving AML Diagnosis

[42] Precision: Deep Learning-Based Classification of Leukocyte Morphology and Malignancy." In 2023 International Conference on Machine Learning and Applications (ICMLA), pp. 2119-2124. IEEE, 2023. https://doi.org/10.1109/ICMLA58977.2023.00320

[43] Mansor, Muhammad Naufal, Mohd Zamri Hasan, Wan Azani Mustafa, Farah Hanan Mohd Faudzi, Syahrul Affandi Saidi, Mohd AminudinJamlos, Noor Anida Abu Talib, and Ahmad Kadri Junoh. "Leukemia Blood Cells Detection using Neural Network Classifier." Journal of Advanced Research in Applied Sciences and Engineering Technology 33, no. 1 (2023): 152-162. https://doi.org/10.37934/araset.33.1.152162

[44] Sakthivel, M., Md Sakiul Islam Sudman, K. Ravishankar, Bhagyalakshmi, Anup Kumar, and Muruganantham. "Medical Image Analysis of Multiple Myeloma Diagnosis Using CNN and KNN Based Approach." In *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pp. 92-97. IEEE, 2023 https://doi.org/10.1109/ICSSAS57918.2023.10331744

[45] Sinaga, 2020, K. P. Sinaga and M.S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796