Leukemia Prediction with AI: A Performance Comparison of K-Means and GMM

Tabasum Guledgudd[1*], Sayed Abulhasan Quadri[2], Noorullah Shariff C[3]

[1] *Department of Computer Science Engineering, A G M Rural College of Enginnering and Technology,Varur, Hubli, India*

[2] *Department of Computer Science Engineering, RIT,Hassan, India*

[3] *Department of AIML, BITM, Ballari, India*

Abstract:

Leukemia, a severe form of blood cancer, disrupts white blood cells, weakening the immune system. Early detection and diagnosis are critical. This study utilizes artificial intelligence, specifically K-means and Gaussian Mixture Model (GMM) clustering algorithms, to classify and predict four types of leukemia—Chronic Myelogenous Leukemia (CML), Acute Lymphocytic Leukemia (ALL), Non-Hodgkin Lymphoma (NHL), and Hodgkin Lymphoma (HL) along with benign cases. Using blood sample reports, we compare the performance of these algorithms based on accuracy, precision, and recall. Accuracy measures the correctness of predictions, precision evaluates the avoidance of false positives, and recall assesses the identification of all leukemia instances. In the course of this study, we have developed a conceptual framework that outlines the end-to-end process, from data acquisition to final classification. This framework integrates clustering techniques to optimize leukemia diagnosis, providing a systematic approach for analysing blood sample data. Our findings highlight the strengths and weaknesses of K-means and GMM, guiding the selection of the most effective algorithm for reliable leukemia diagnosis through AI and machine learning techniques.

Keywords: AIML, K-means, Gaussian Mixture Model, Classification, Clustering, leukemia, Metrics.

Introduction

Cancer results from cells growing uncontrollably due to genetic mutations, leading to more cancerous cells. Leukemia, a type of blood cancer, impacts white blood cells (WBC) and weakens the immune system, hindering its ability to combat diseases and viruses. This cancer can be acute, with rapid growth, or chronic, with slower progression, and it targets either lymphoid cells or myelocytes (young granulocyte cells). Monitoring WBC levels is vital for early detection and maintaining a strong immune system (Blackadar 2016).

Artificial intelligence in healthcare uses machine-learning algorithms to mimic or surpass human cognition in analysing and understanding complex medical data, aiding in diagnosis, treatment, and disease prevention. Numerous diagnostic techniques are used to identify different types of leukemia, comprising blood tests, bone marrow aspiration and biopsy, cytogenetic analysis, molecular testing, flow cytometry, and imaging tests. This work introduces a scientific approach for detecting leukemia using blood sample reports from infected individuals.

Among the numerous AIML algorithms available, we opted for the good old primitive K-means and GMM, as they enable an unbiased, data-driven exploration of leukemia subtypes without requiring pre-labelled data. This ensures greater scalability, adaptability, and real-world applicability, particularly in cases where supervised learning models are constrained by the availability and quality of labelled datasets.

Two clustering algorithms are utilized to classify and predict four types of leukemia: CML, ALL, NHL, HL along with benign cases.

One of the major parts of the work is based on comparison of these algorithms using performance metrics. Comparing K-means and GMM across accuracy, precision, and recall showcase insights into their performance in predicting leukemia types. Accuracy highlights which algorithm better predicts correct types with fewer misclassifications. Precision assesses how well each algorithm avoids false positives, ensuring precise predictions. Recall evaluates each algorithm's ability to identify all leukemia instances, minimizing false negatives. These comparisons guide the selection of the most effective algorithm for accurate and reliable leukemia diagnosis using AI techniques.

The paper is organized as follows: it begins with a discussion on clustering technology, followed by a review of the relevant literature. The methodology is then presented, after which the results are discussed. Finally, the paper concludes with a summary of findings and future directions.

Clustering Technology

Clustering involves dividing data into groups based on similarities, using metrics like Euclidean distance or cosine similarity. Key elements include clusters, centroids (or medoids), and distance metrics, pivotal in grouping similar data points. Applications include unsupervised learning, customer segmentation, anomaly detection, image and text clustering, personalized recommendations, genomics, spatial analysis, and more. Common algorithms like K-means, GMM, Hierarchical Clustering, DBSCAN, and Mean Shift offer diverse approaches.

Many clinical studies still rely on clustering for disease subtyping, risk stratification, and anomaly detection. Using well-established clustering techniques like K-means and GMM ensures interpretability in medical decision-making (Shah et al. 2021; Tabasum et al. 2024).

We explore K-means and GMM effectiveness in leukemia diagnosis, emphasizing research contributions. Comparatively, K-means and KNN share similarities, pivotal in clustering approaches.

K-means

K-means Clustering is an unsupervised learning algorithm widely used in machine learning and data science for clustering tasks. It groups an unlabelled dataset into K clusters based on similarity. The parameter K specifies the number of clusters to be created; for example, K=2 results in two clusters, and K=3 results in three clusters. This iterative algorithm allocates each data point to the cluster whose centroid is closest, aiming to minimize the distance between data points and their respective centroids. K-means is effective for identifying inherent groupings within data sets without prior training, functioning on the principle of associating each cluster with its centroids.

GMM

Gaussian Mixture Models (GMMs) expand upon k-means clustering by fitting multi-dimensional Gaussian distributions to datasets. They provide a probabilistic measure of cluster assignment certainty and accommodate complex cluster shapes. Unlike k-means, GMMs employ a generative probabilistic model to describe data distribution and determine optimal components using likelihood evaluation and cross-validation to prevent over-fitting.

K-means and KNN: Uncovering the Shared Principles

Despite some differences, K-means and KNN share many similarities, both being crucial for data analysis and modelling. They are distance-based approaches, computing distances between data points in a multi-dimensional feature space. K-means uses Euclidean distance to assign points to the nearest cluster centroid, refining clusters to minimize within-cluster

variance. KNN uses distance metrics to find the K nearest neighbours for classification or regression.

Both algorithms treat data points as vectors in high-dimensional space, relying on distance metrics, making feature scaling important. Initial parameters, such as the number of clusters in K-means or the number of neighbours in KNN, must be predefined, influencing their performance. The choice of K is often guided by domain knowledge or validation techniques like cross-validation.

K-means and K-nearest neighbours (KNN) are non-parametric algorithms, adaptable to diverse datasets without assuming underlying data distributions. K-means offers interpretable results through cluster visualization, while KNN predicts based on proximity to labelled neighbours.

Both algorithms are straightforward to implement but vary in computational workload. K-means complexity scales with data volume, cluster count, iterations, and dimensions, whereas KNN's computation primarily depends on data size and dimensions during distance calculations.

Literature review

This section explores prior research where researchers have employed K-means and GMM clustering techniques for classification and prediction. The literature review encompasses studies by other researchers, focusing on evaluating these algorithms for comparison.

Advancements in technology and space exploration have made it possible to collect vast spatial-temporal data. Analysing and deriving valuable insights from this data using data mining techniques, especially clustering methods, is essential for addressing real-world challenges. Herein, we present previous relevant research conducted in this domain.

Adinanta et al. (2020) proposed using computer vision techniques to detect individuals and assess physical distancing violations. Unlike traditional GPU-intensive object detection methods, this work employs background subtraction methods based on Gaussian Mixture Models (GMM), such as GMG, KNN, MOG, and MOG2, which require less computation. Performance evaluation shows that KNN is the best method, followed by MOG, MOG2, and GMG in terms of the various metrics and detection accuracy. This study demonstrates the effectiveness of background subtraction techniques for the efficient social distancing monitoring.

Sammali et al. (2021) introduced a novel approach using multi-modal uterine-activity measurements to predict embryo implantation success in IVF. The study integrated features

from electro hysterography (EHG) and B-mode transvaginal ultrasound (TVUS) recordings. Machine learning models including KNN and GMM classified uterine activity as favourable or adverse for implantation. KNN achieved the highest accuracy of 93.8% across all phases (follicular stimulation, ET1, and five to seven days post-ET), highlighting its predictive capability. The study emphasizes multi-parametric strategies to enhance IVF success rates through objective assessment of uterine receptivity and embryo quality.

Singh et al. (2022) focused on diagnosing Hepatitis C Virus (HCV) using dual datasets, highlighting the need for accurate diagnosis to improve treatment outcomes. The study used supervised learning models (Decision Tree, Logistic Regression, KNN) for classification and unsupervised learning models (K-means, Hierarchical Clustering, DBSCAN, Gaussian Mixture) for clustering HCV data. Logistic Regression excelled in classification, while K-means performed best in clustering. These findings demonstrate machine learning's potential to enhance diagnostic accuracy and treatment effectiveness for Hepatitis C.

Hemachandira and Viswanathan (2022) focused on detecting epileptic seizures from EEG signals using mathematical model-based classifiers and efficient feature selection. They employed wavelet transforms (Haar, dB4, Sym 8) on the Bonn epilepsy dataset and optimized features with Particle Swarm Optimization (PSO). The SVM classifier with RBF kernel and Sym 8 wavelet features selected by PSO achieved 98% accuracy with a 2% error rate, outperforming other classifiers. Early diagnosis supports patient rehabilitation, suggesting further exploration into deep neural networks and advanced classifiers.

Chen et al. (2022) used dictionary learning to create an over complete basis for sparsely representing geochemical exploration data in Chengde, Hebei Province, China. They analyzed gold mineral anomalies using five algorithms and assessed anomaly levels via Euclidean distance from sparse representations. These models surpassed traditional methods like KNN and GMM in ROC curve and AUC metrics, successfully identifying prospective gold areas that correlate with known deposits and geological indicators. The study underscores dictionary learning's effectiveness in mineral exploration and suggests further validation across diverse geological settings.

Epilepsy diagnosis relies on accurate EEG analysis to capture neuronal activity. Hemachandira et al. (2022) employed wavelet transforms (Haar, dB4, Sym 8) on the Bonn epilepsy dataset, optimizing feature selection with Particle Swarm Optimization (PSO). Their study evaluated seven classifiers, with SVM using RBF kernel achieving 98% accuracy and a 2% error rate. This highlights the effectiveness of wavelet-based feature extraction and PSO-driven selection in enhancing EEG-based epilepsy detection.

Kalaiyarasi and Harikumar (2022) proposed ovarian cancer detection using microarray gene data, crucial for accurate diagnosis. They employed ANOVA for gene selection and clustering-based (FCM) and transform-based (SDA, Hilbert, FFT, DCT) methods for feature extraction. Classifiers achieved varied accuracies, with the NLR classifier performing best using SDA features (92% accuracy) and GMM achieving 88% accuracy with correlation distance feature selection. Challenges include dataset complexity and model constraints, suggesting future research in heuristic feature selection for improved classification.

Rasul et al. (2022) assessed machine learning models and clustering algorithms for detecting autism spectrum disorder (ASD) across diverse datasets. SVM and Logistic Regression showed strong performance for children, while Logistic Regression performed best for adults. The ANN model demonstrated superior performance across combined datasets. Spectral clustering effectively grouped ASD-related traits. Key findings included identifying significant ASD characteristics like A9 (aversion to physical contact in adults) and A4 (challenges in understanding others' emotions in children). Future research aims to apply advanced models to larger datasets and integrate deep learning for improved feature learning and classification. The proposed ANN model achieved 94.24% accuracy for the combined dataset after hyperparameter tuning. Spectral clustering outperformed other methods in clustering accuracy metrics. Code implementation is available on GitHub.

Wang et al. (2023) introduced a novel clustering algorithm combining Gaussian Mixture Models (GMM) and K-Nearest Neighbours (KNN) to address temporal discontinuity in clusters. Tested on a spatial-temporal temperature dataset across 42 locations over a year, the algorithm maintained temporal continuity within clusters. Application to different time points and years validated its effectiveness in clustering spatial-temporal data based on temperature, demonstrating transferability and real-world applicability.

Haibin et al. (2023) introduced a method to mitigate classifier bias in imbalanced datasets using a borderline oversampling approach. They combined K-nearest neighbour (KNN) and Deep Gaussian Mixture Model (DGMM) to identify and oversample minority samples. This method improved AUC ROC Curve and G-mean by up to 8.62% and 12.99%, respectively, with average increases of 3.51% and 4.93%. The authors noted the need for further optimization of DGMM parameters, but overall, the approach effectively enhances classifier performance in handling imbalanced data challenges.

Pandya et al. (2023) explored epilepsy detection using advanced machine learning techniques. They analyzed EEG data from the Children's Hospital Boston and Massachusetts Institute of Technology database, introducing novel features like root mean square of RMS,

mean absolute value, and waveform length (RRWM) alongside the generalized method of moments (GMM). The K-nearest neighbour (KNN) algorithm with a window size of 300 and hyper-parameter tuning emerged as the most effective model, achieving 99.03% accuracy with a time complexity of 13 milliseconds. GMM exhibited the fastest computation time of 8.99 milliseconds for epilepsy detection. This study underscores the role of novel features and machine learning in enhancing epilepsy prediction accuracy and computational efficiency.

Lin et al. (2023) developed a highly accurate indoor localization method for 4G and 5G-equipped buildings. They introduced a two-stage clustering-based approach (TSCA), combining a novel group matching method with KNN variants. TSCA significantly outperformed traditional KNN methods, achieving a 55% improvement in positioning accuracy. Specifically, it showed a 13.36% improvement in 2D accuracy and 10.3% in 3D accuracy using Wi-Fi Received Signal Strength fingerprints.

Kalaiyarasi et al. (2023) proposed using Short-Time Fourier Transform (STFT) on Photoplethysmography (PPG) signals to detect respiratory disorders non-invasively. The method monitors respiratory patterns effectively, achieving 78.05% accuracy with Bayesian Linear Discriminant Classifier (LDC). This approach enhances diagnostic capabilities for early intervention in respiratory health and improves reliability in myoelectric control systems by rejecting unreliable movements, surpassing traditional classifiers like LDA, GMM, and KNN.

Archana and Kumar (2023) proposed using GLCM to extract six texture features (Contrast, Correlation, Energy, Homogeneity, Dissimilarity, and ASM) from Dermnet image library images. Gaussian Mixture Model (GMM) approximations were applied to each feature's distribution during dataset creation. They evaluated classification with K Nearest Neighbour (KNN), Support Vector Machine (SVM), and Logistic Regression, finding KNN with Cosine Similarity outperformed SVM and Logistic Regression in accuracy. Their approach improves automated skin disease classification systems, aiding early detection and treatment.

Abid et al. (2023) introduced Smart K Nearest Neighbor Outlier Detector (SKOD) to improve EEG signal quality in e-health applications by mitigating artifacts. SKOD, a non-parametric, unsupervised algorithm, uses Euclidean distances to identify outliers in EEG data without initial KNN parameter configuration. Evaluated on BCI competition II benchmark data, SKOD achieved robust performance with sensitivity and specificity exceeding 60%,

demonstrating nearly perfect outlier detection rates and promising applications for real-time artefact management.

Table 1: Critical Analysis of Literature on Spatial-Temporal Data and Machine Learning Applications

Author(s)	Year	Methodology/ Technique Used	Dataset/Input Used	Results (Accuracy, Precision, Recall)	Performance	Limitations
Adinanta et al.	2020	Background subtraction (GMM, GMG, KNN, MOG, MOG2)	Physical distancing monitoring dataset	KNN performed best among methods	Effective social distancing monitoring with low computation	Limited to physical distancing monitoring, not generalizable
Sammali et al.	2021	Multi-modal uterine-activity measurements using SVM, KNN, GMM	Electrohysterography (EHG) & B-mode transvaginal ultrasound	KNN achieved 93.8% accuracy	Enhanced IVF success prediction	Requires further validation on larger datasets
Singh et al.	2022	Supervised (Decision Tree, Logistic Regression, KNN) & Unsupervised (K-means, Hierarchical, DBSCAN, GMM) learning	Dual dataset for Hepatitis C Virus (HCV) diagnosis	Logistic Regression excelled in classification, K-means best in clustering	Improved diagnostic accuracy	Limited exploration of deep learning models
Chen et al.	2022	Dictionary learning for geochemical anomaly detection	Chengde, Hebei Province geochemical data	Outperformed KNN, GMM in AUC, ROC metrics	Successfully identified mineral deposits	Requires validation across diverse geology
Kalaiyarasi &	2022	ANOVA for gene selection,	Microarray gene data for ovarian	NLR classifier	Accurate classification	Dataset complexity

Harikumar		FCM, SDA, Hilbert, FFT, DCT for feature extraction	cancer detection	(92% accuracy), GMM (88%)	of ovarian cancer	and model constraints
Wang et al.	2023	Gaussian Mixture Models (GMM) and K-Nearest Neighbors (KNN) for spatial-temporal clustering	Temperature dataset (42 locations, 1 year)	Maintained temporal continuity in clusters	Transferable across different datasets	Needs testing on non-temperature datasets
Haibin et al.	2023	Borderline oversampling with KNN, Deep Gaussian Mixture Model (DGMM)	Imbalanced datasets	Improved AUC ROC (+8.62%), G-mean (+12.99%)	Enhanced classifier performance for imbalanced data	Requires DGMM parameter optimization
Lin et al.	2023	Two-stage clustering-based approach (TSCA) with KNN variants	Wi-Fi Received Signal Strength fingerprints (4G/5G buildings)	55% improvement in positioning accuracy	Enhanced indoor localization	Requires real-world deployment testing
Kalaiyarasi et al.	2023	Short-Time Fourier Transform (STFT) on PPG signals	PPG dataset for respiratory disorder detection	Bayesian LDC achieved 78.05% accuracy	Non-invasive respiratory disorder detection	Performance lower than deep learning approaches
Archana & Kumar	2023	GLCM texture feature extraction, GMM approximations, KNN, SVM, Logistic Regression	Dermnet image library	KNN with Cosine Similarity outperformed SVM, Logistic Regression	Improved skin disease classification	Needs validation on larger dermatological datasets
Abid et al.	2023	Smart KNN Outlier Detector	BCI competition II dataset	Sensitivity, specificity exceeded	Improved real-time EEG artifact	Requires higher sensitivity

		(SKOD) for EEG signal quality		60%	management	and specificity

This structured chart organizes the literature review in an analytical manner, making it easier to compare methodologies, datasets, results, and limitations.

In our research, we aim to provide a comparative analysis of two clustering algorithms. Clustering Methods do not require large datasets, K-Means and GMM work efficiently with moderate dataset sizes, unlike deep learning models, which require massive labelled data.
While each algorithm demonstrates strengths in different contexts, determining a definitive superior is challenging due to their distinct potentials. Our study will assess these algorithms based on fundamental metrics that include accuracy, precision, and recall.

Methodology

Overview of system implementation

In this research, we propose a conceptual framework that outlines the complete workflow of leukemia diagnosis using clustering techniques. This framework, formulated through our analytical approach, systematically defines each step from data acquisition to final classification.

Figure 1 illustrates the complete workflow of the proposed AI-ML-based leukemia classification framework. The process begins with the Complete Blood Count (CBC) Report, followed by data acquisition to collect relevant medical data. Once the data is acquired, it undergoes a pre-processing stage, which includes checking for valid data to ensure completeness, identifying and handling missing values, and organizing the dataset for further analysis. After pre-processing, the processed data is fed into an AI/ML implementation stage, where unsupervised learning algorithms are applied. Specifically, two clustering methods, K-means and the Gaussian Mixture Model (GMM), are employed for analysis. Based on the clustering results, the system classifies leukemia into different types, including Chronic Myelogenous Leukemia (CML), Acute Lymphocytic Leukemia (ALL), Non-Hodgkin Lymphoma (NHL), Hodgkin Lymphoma (HL), and benign cases (non-leukemic cases). This structured framework enables efficient leukemia classification, assisting in early detection and improving diagnostic accuracy using AI and machine learning techniques. The framework ensures a structured approach to leukemic detection by integrating clustering-

based machine learning methods, ultimately aiding in selecting the most effective algorithm for accurate and reliable diagnosis.

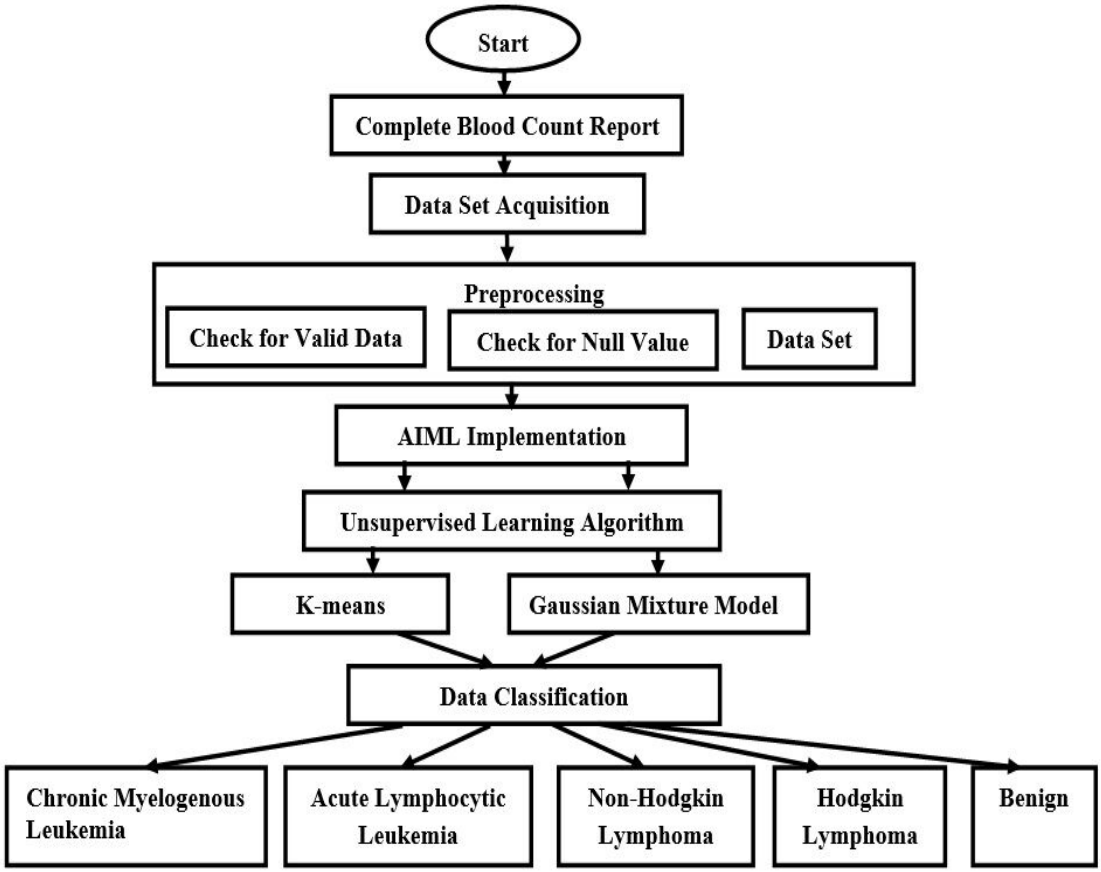


Figure 1: Proposed framework for leukemia diagnosis and classification

Overview of system implementation is shown in the Figures 2 and Figure 4 The flowcharts outline 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 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 leukemia into four types. The same procedure is also applied to GMM as shown in the Figure 4 The validation of AIML techniques involves applying diverse performance metrics to assess both algorithms.

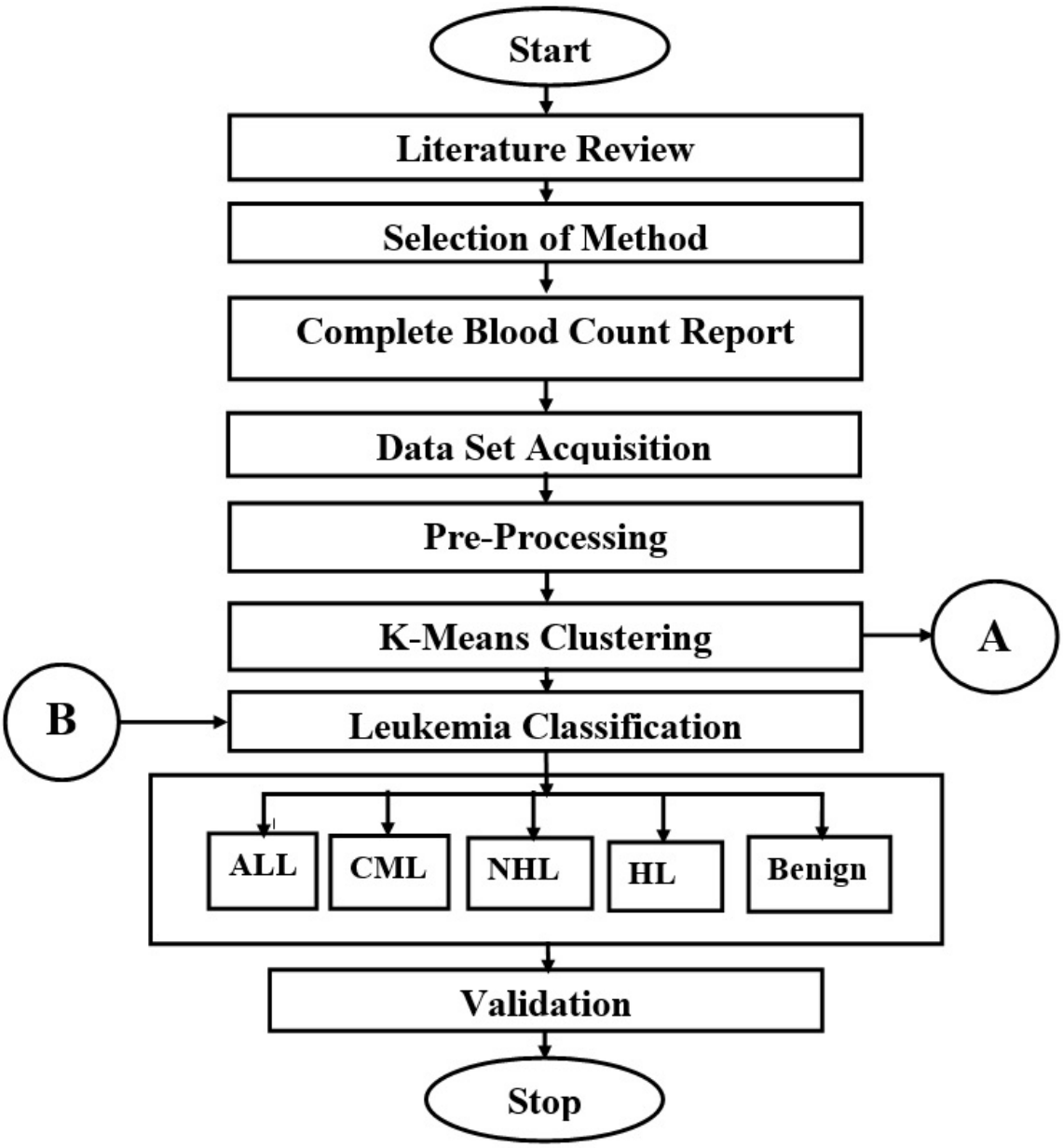


Figure 2: System implementation flowchart (using K- means)

K-means clustering is a machine learning algorithm that partitions data points into 'K' clusters based on similarities in their features. Initially, 'K' centroids are randomly placed, and data points are assigned to the nearest centroid. Centroids are then recalculated as the mean of assigned data points, iteratively optimizing cluster compactness until convergence, where assignments stabilize. The algorithm aims to minimize within-cluster sum of squares to enhance cluster separation and coherence. Figure 3 illustrates the steps of the K-means clustering process.

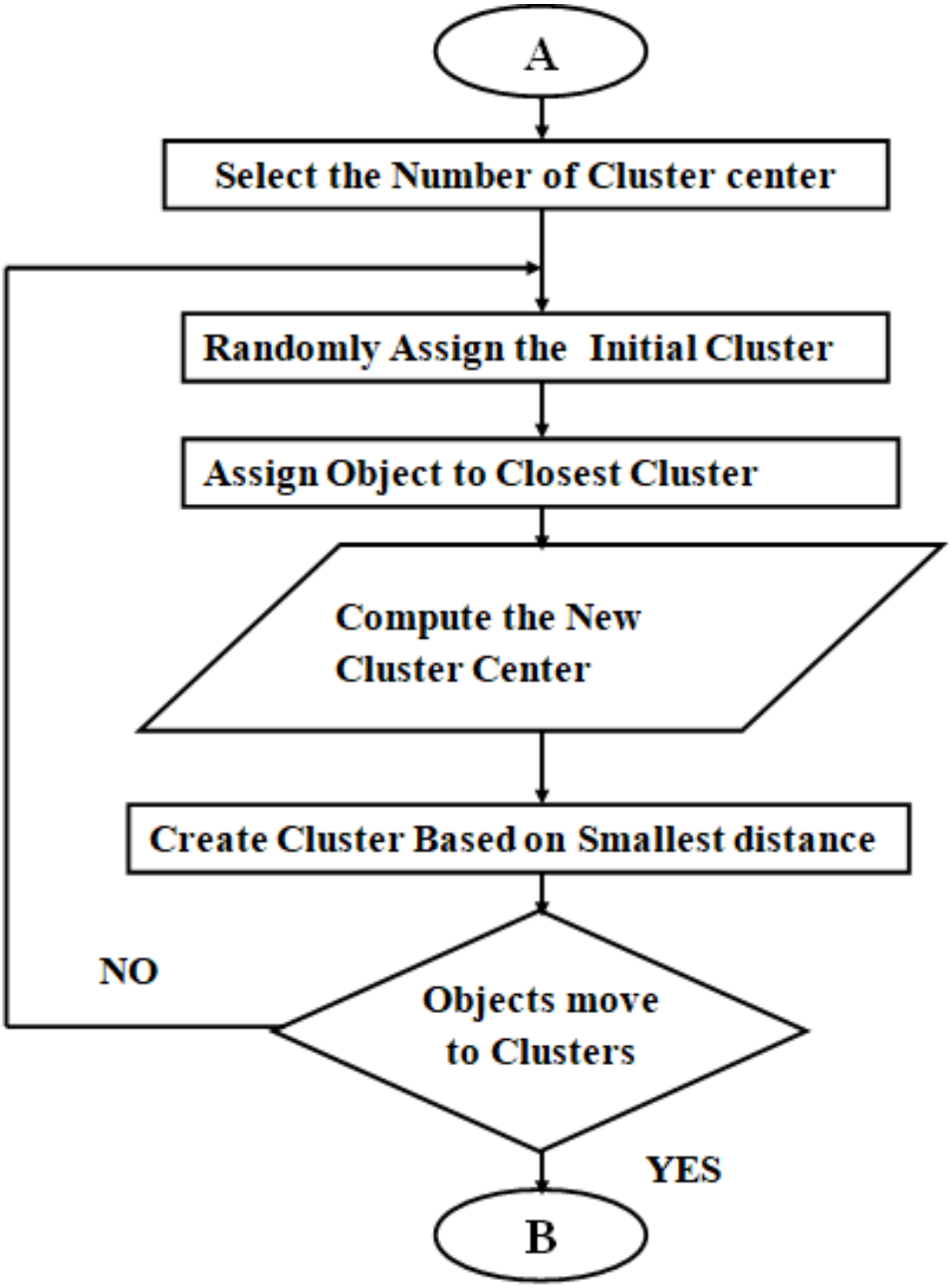


Figure 3: System Implementation Flowchart (K- means working)

The procedure for GMM is conferred in the following section.

As shown in the figure 5 GMM algorithm starts by inputting a sample set with Gaussian numbers and initializing the model parameters. It calculates the posterior probability of each component to determine the likelihood of data points belonging to each Gaussian distribution. Using these probabilities, the algorithm iteratively updates the mean vector, covariance matrix, and mixing coefficients, refining the model parameters. A key step in GMM is the convergence check, where the algorithm assesses if the parameters have stabilized. If not, it

repeats the calculations and updates until convergence is achieved. Once the model converges, the parameters are finalized to best fit the data.

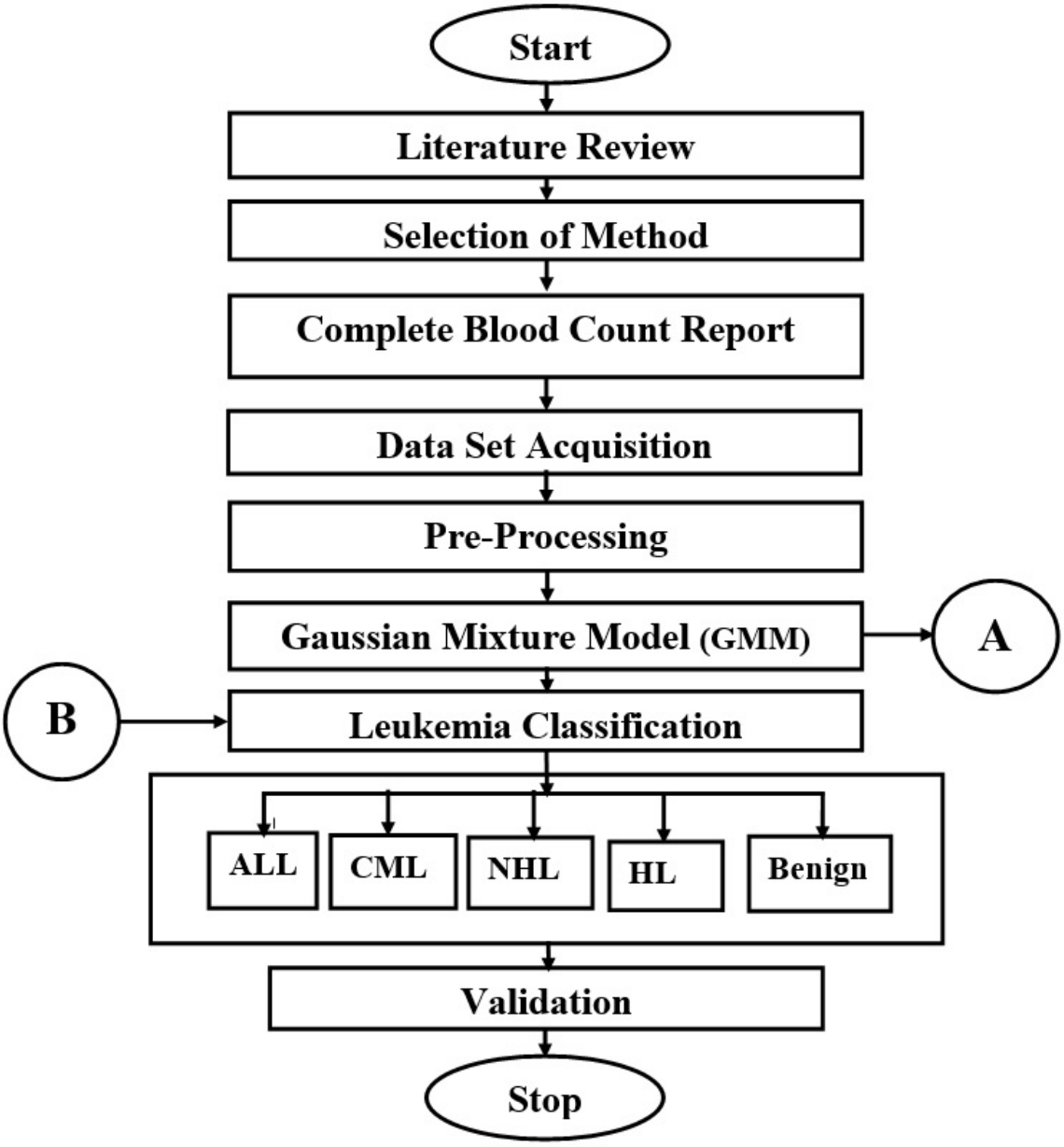


Figure 4: System implementation flowchart (using GMM)

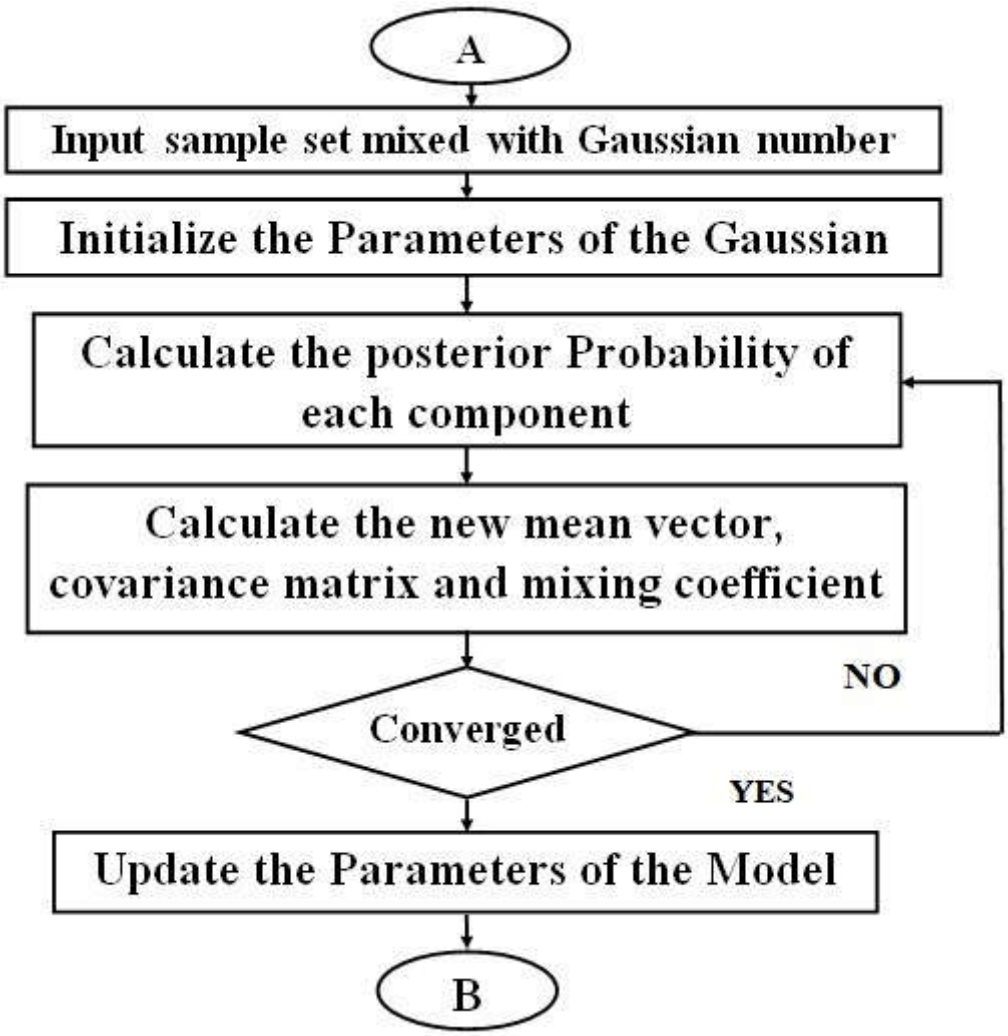


Figure 5: System Implementation Flowchart (Gaussian Mixture Model working)

In this work, 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.

Data acquisition

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/).

The dataset consists of real-time patient CBC reports directly collected from a cancer hospital, making it subject to ethical, privacy, and logistical constraints. Unlike publicly available datasets, access to medical data is limited due to strict regulatory compliance.

Reports of 300 Patients diagnosed with five distinct types of blood cancer were gathered for the study. 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.

Data Pre-processing

Pre-processing is crucial for providing valid data inputs essential for an effective clustering process. Its significance lies in its impact on the accuracy of the results. One key aspect involves handling missing data and addressing any null values present in the dataset. Given our dataset comprises 13 parameters from 300 samples, we undertake manual inspection to identify and rectify any irrelevant or null data entries, ensuring the integrity of the dataset before proceeding with clustering.

Conducting K-means clustering

The training phase entails employing the K-means algorithm to extract patterns from the dataset. The model is trained to distinguish between clusters corresponding to benign and blood cancer conditions. Execution of the program is conducted within the MATLAB IDE, utilizing the K-means Algorithm. Visualizing the clustering outcomes aids in comprehending how data points are grouped or clustered based on specific similarities or features. Figure 6 illustrates the visual scatter output of the K-means clustering results.

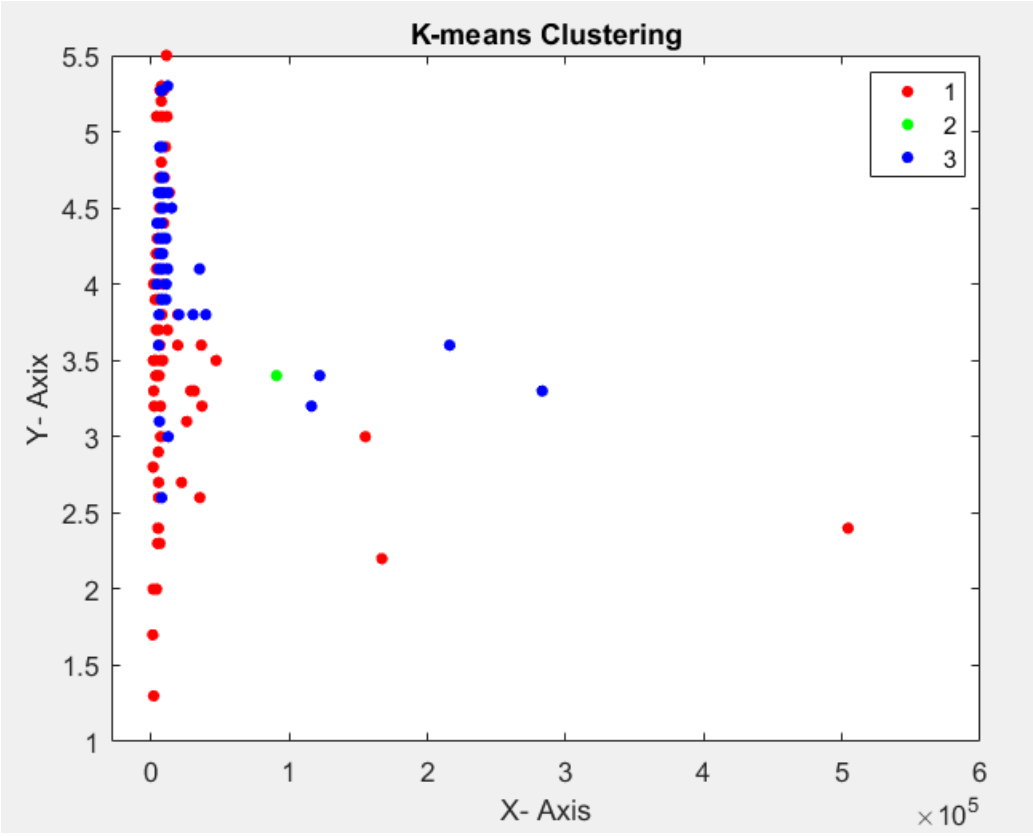


Figure 6: K-means Clustering

Data is visualized using data points and cluster centroids is shown in Figure 7

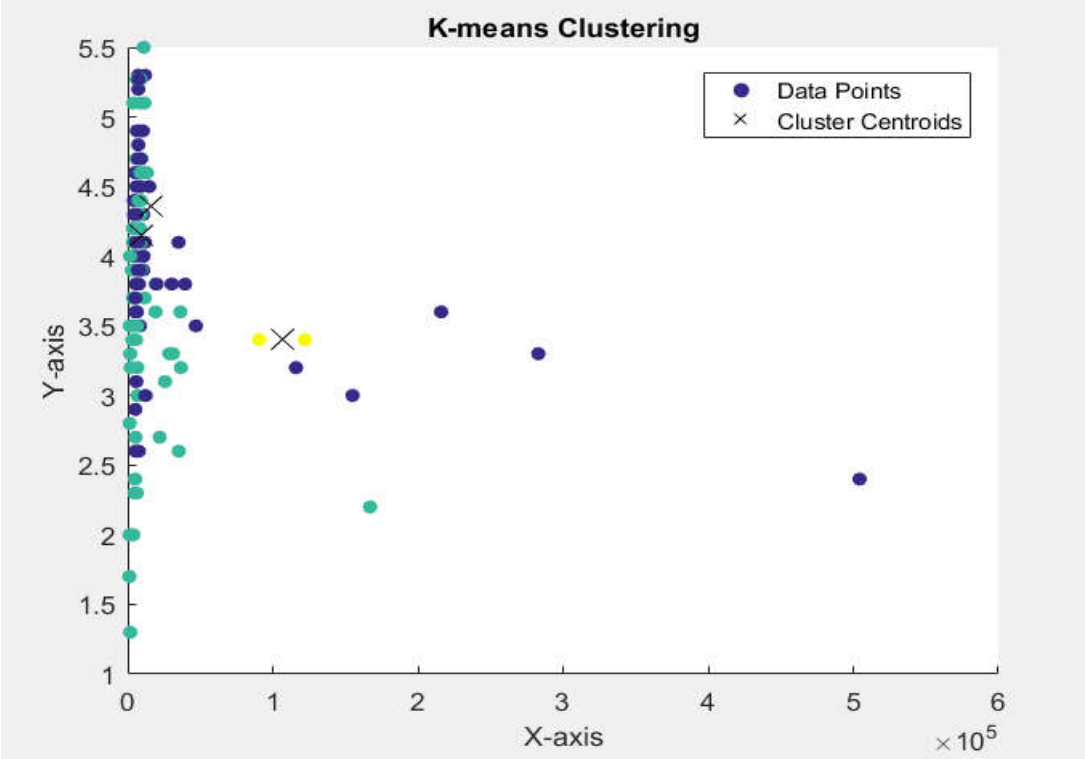


Figure 7: Data points versus cluster centroids – K means Clustering

Understanding the roles of data points and cluster centroids is fundamental in clustering algorithms like K-means. Data points are individual observations described by multiple features, forming the basis of the dataset. Each data point is represented as a vector in the feature space. Cluster centroids, on the other hand, represent the centre of a cluster and are computed as the mean of all data points in that cluster. They guide cluster formation by minimizing distances between data points and centroids. This understanding is crucial for applications such as pattern recognition and data mining, where clustering algorithms organize and interpret complex datasets based on similarities among data points.

Conducting GMM clustering

GMM offers a probabilistic approach to clustering that differs from methods like K-means. GMMs assume that the data points are generated from a mixture of several Gaussian distributions, each representing a cluster. This model provides a flexible means of capturing the inherent structure of the data by allowing clusters to have different shapes, sizes, and orientations.

GMM clustering is shown in figure 8. In GMMs, data points represent individual observations within a dataset, much like in K-means. Each data point is characterized by multiple features or attributes and is depicted as a vector in the feature space. Data points versus cluster centroids are visualized in figure 9. The goal of GMMs is to model the probability distribution of these data points by fitting a mixture of Gaussian distributions.

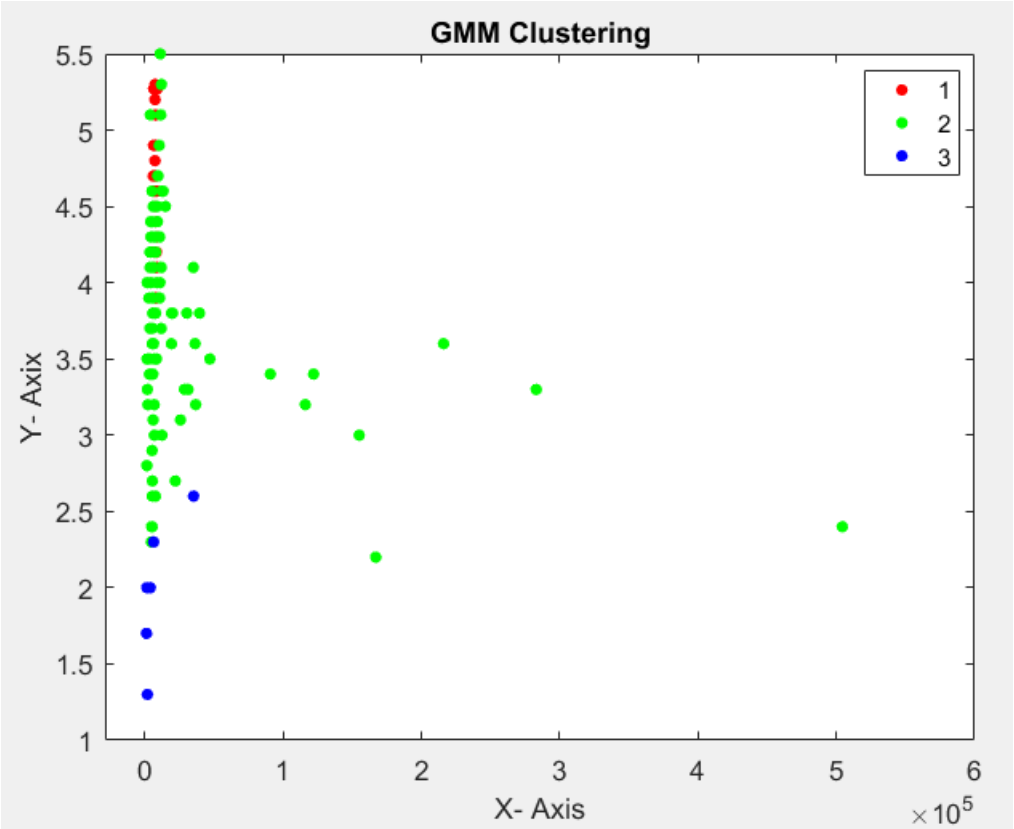


Figure 8: GMM Clustering

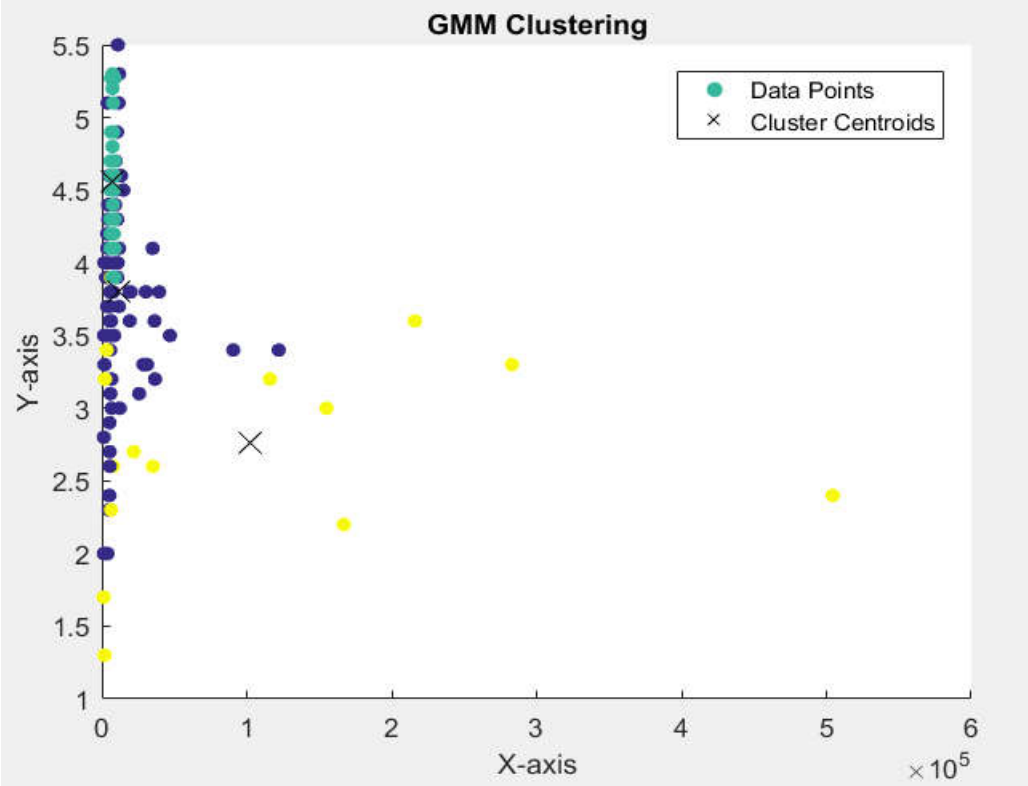


Figure 9: Data points versus cluster centroids – GMM Clustering

Prediction

Absolutely, once new values for any of the thirteen blood parameters are inputted, the trained model utilizes clustering to assign these values to the nearest cluster. Based on this assignment, the model predicts whether the patient's condition leans towards benign or indicates potential blood cancer. If cancer is suggested, the model further predicts its specific class. This process provides insights into disease likelihood and classification, supporting diagnosis and treatment decisions.

Results and discussions

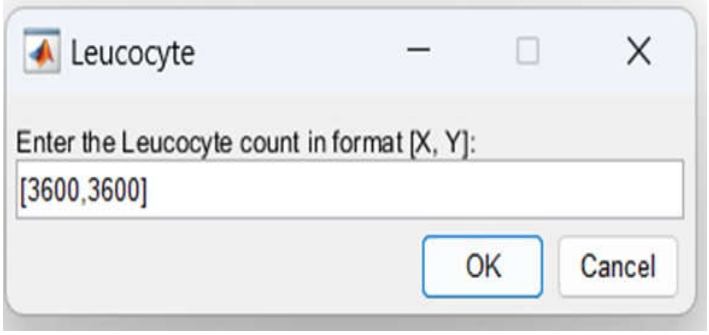
Findings and Analysis

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 four types of Leukemia along with benign cases. Below are tables displaying five significant blood parameters:

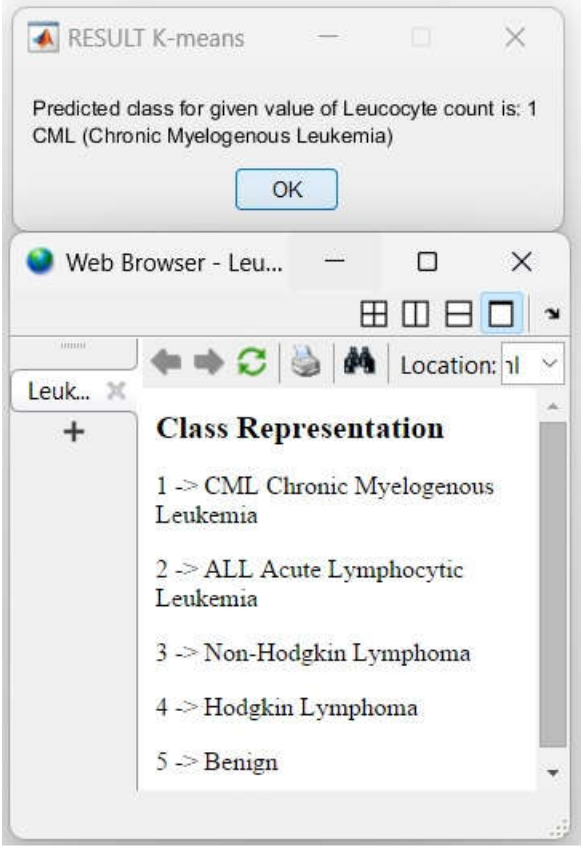
Table 2: Hematological Parameters of Patient-1

(Input for Unsupervised Learning Algorithm)

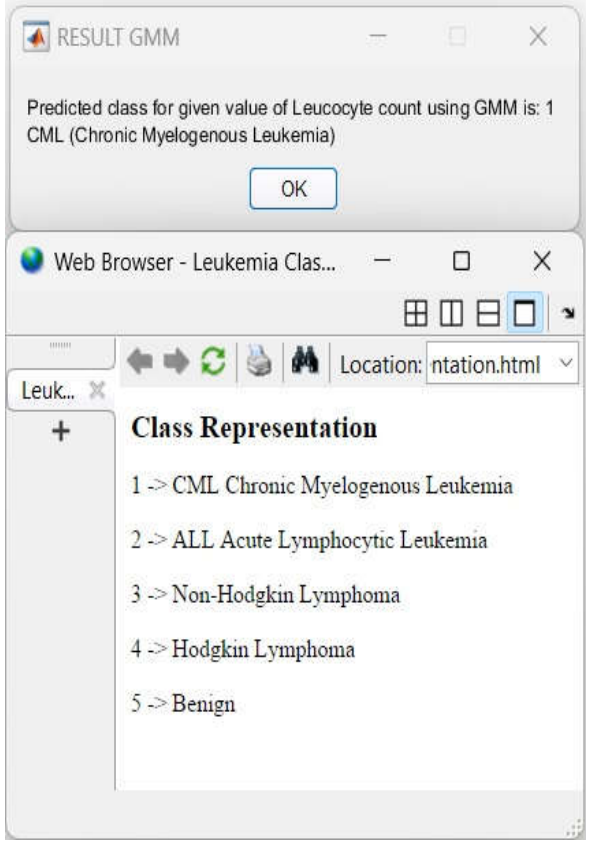
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
3600	40	31	2	1	Chronic Myeloid Leukemia



Input value of Leucocyte count (Patient-1)

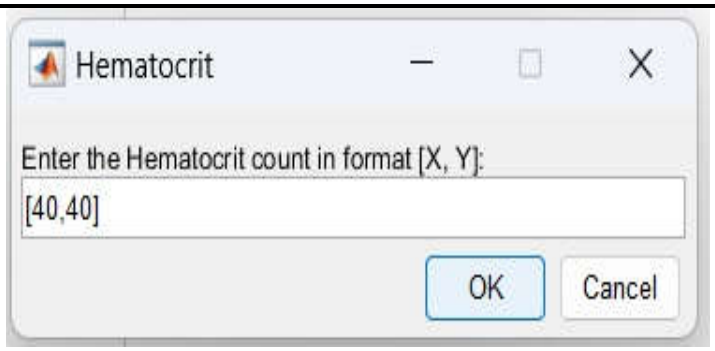


Output predicted class for patient-1 using K-means

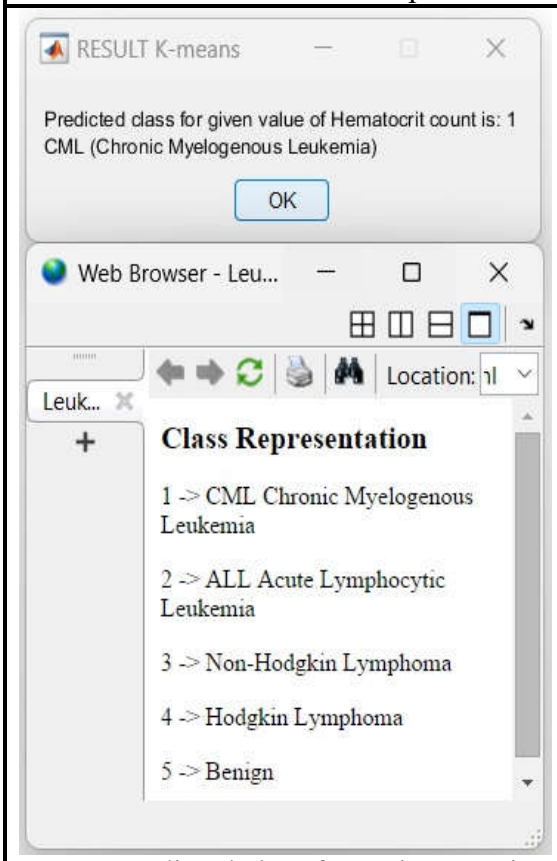


Output predicted class for patient-1 using GMM

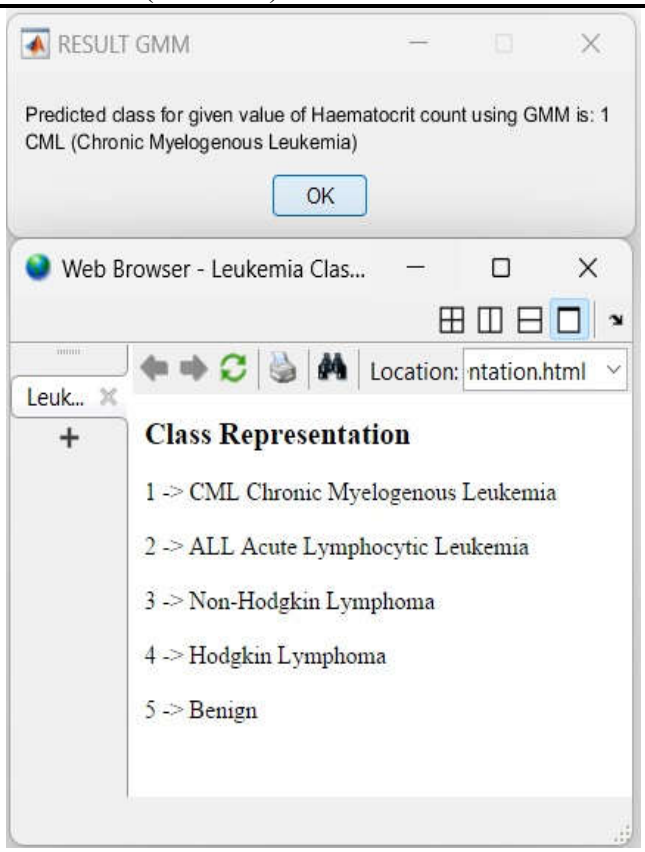
Figure 10: Input and Output of Leucocyte count (Patient-1)



Input value of Hematocrit (Patient-1)

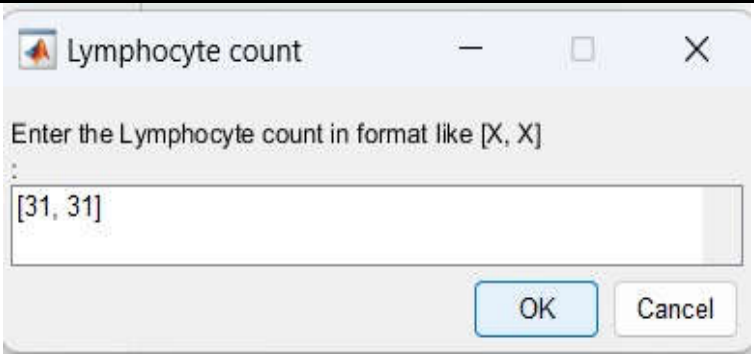


Output predicted class for patient-1 using K-means

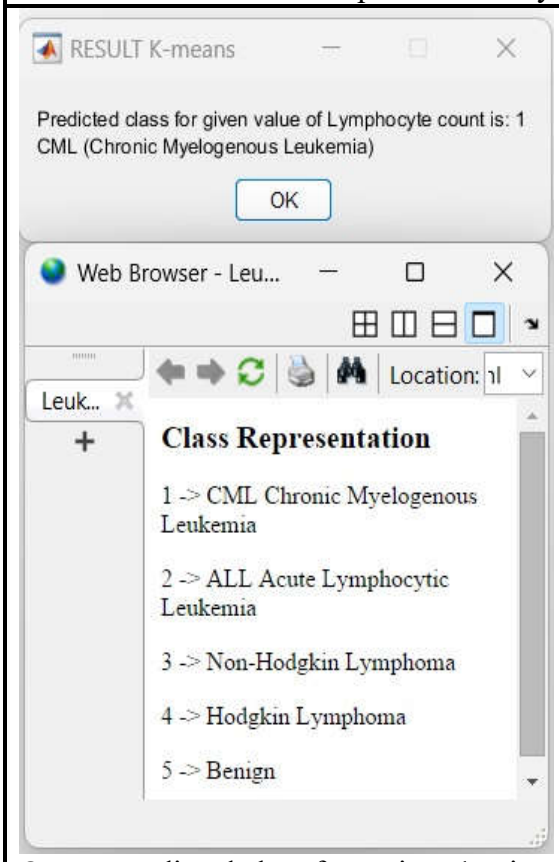


Output predicted class for patient-1 using GMM

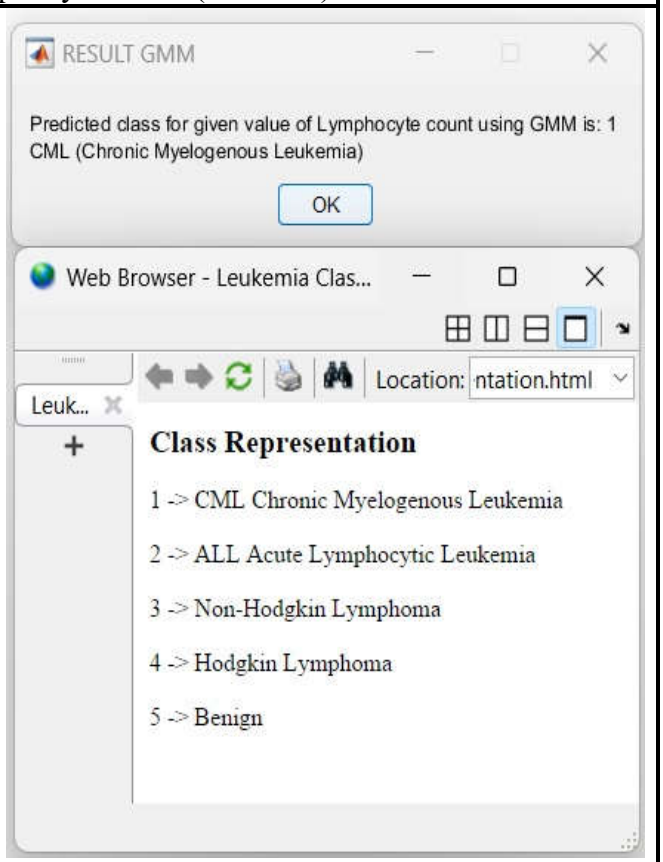
Figure 11: Input and Output of Hematocrit count (Patient-1)



Input value of Lymphocyte Count (Patient-1)

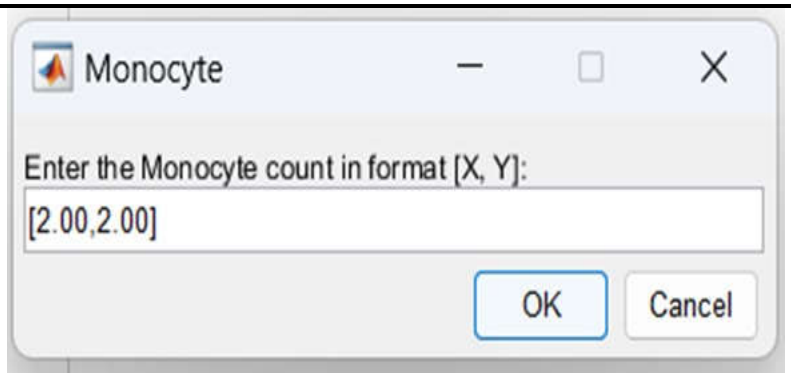


Output predicted class for patient-1 using K-means

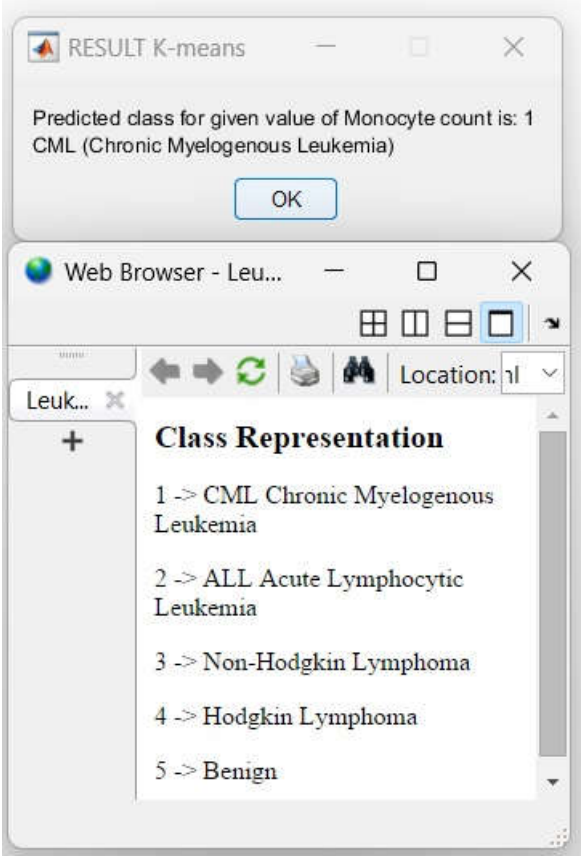


Output predicted class for patient-1 using GMM

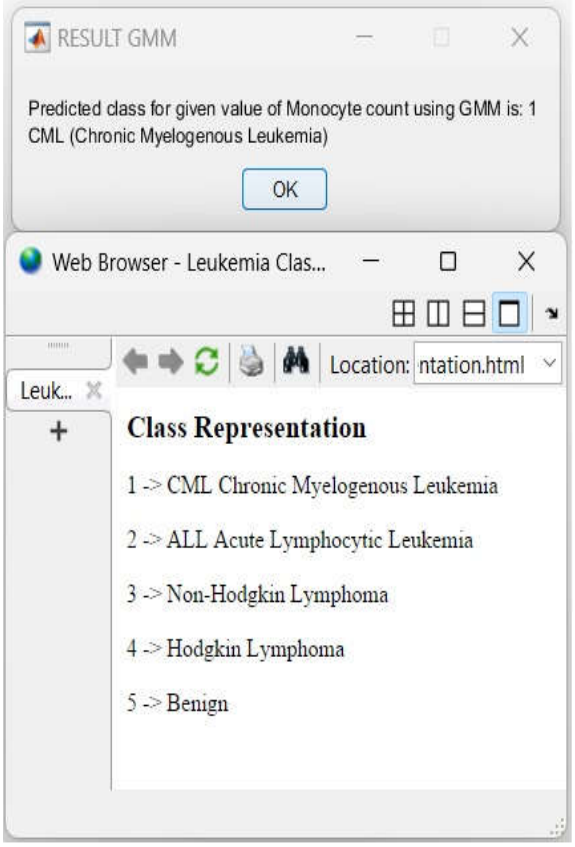
Figure 12: Input and Output of Lymphocyte count (Patient-1)



Input value of Monocyte (Patient-1)

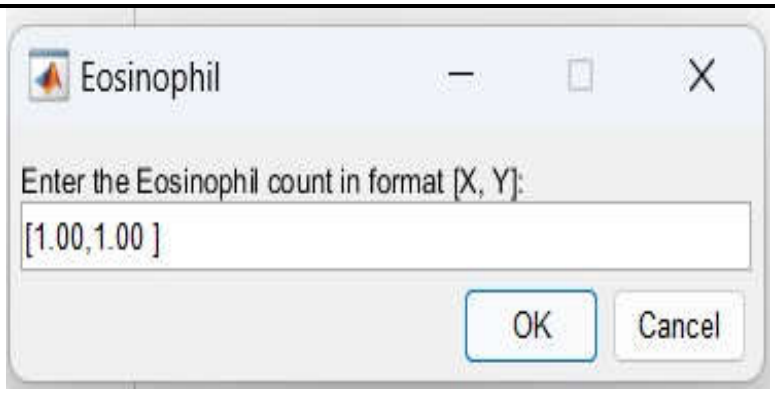


Output predicted class for patient-1 using K-means

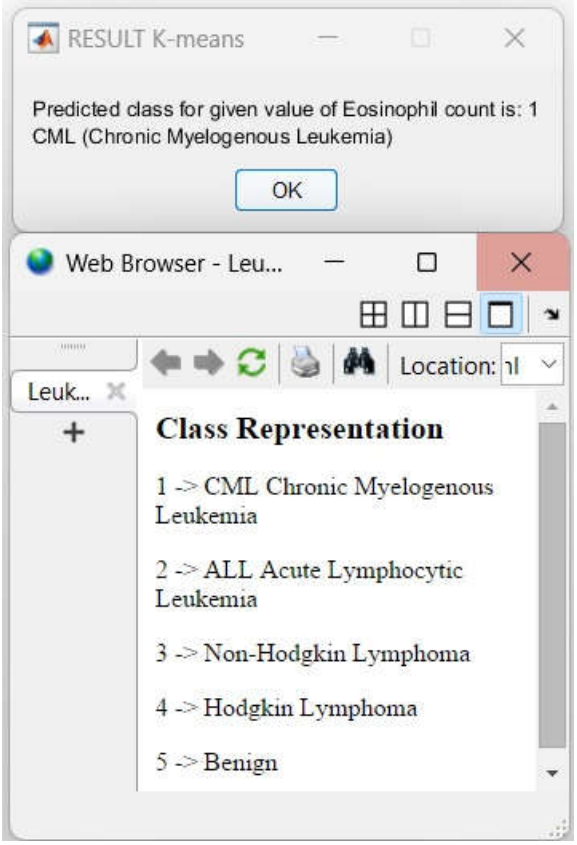


Output predicted class for patient-1 using GMM

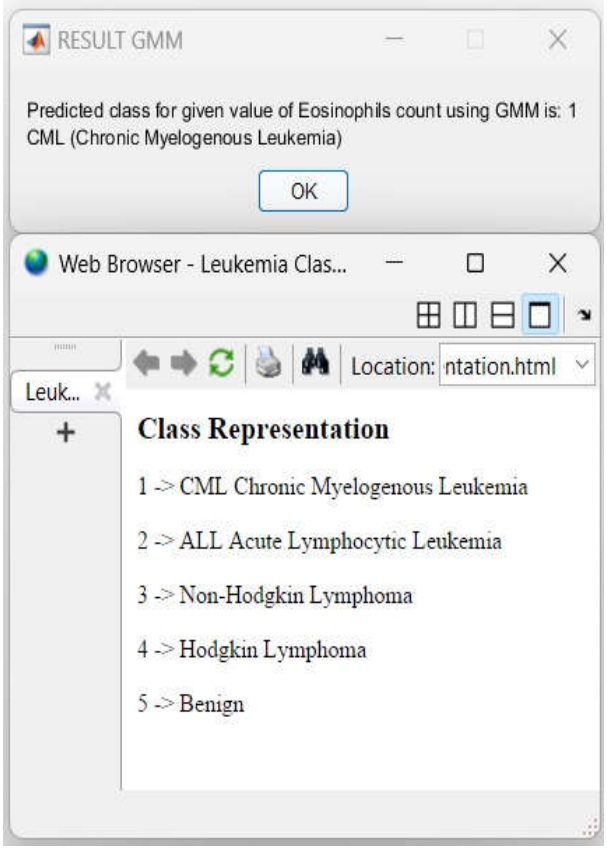
Figure 13: Input and Output of Monocyte count (Patient-1)



Input value of Eosinophil (Patient-1)



Output predicted class for patient-1 using K-means



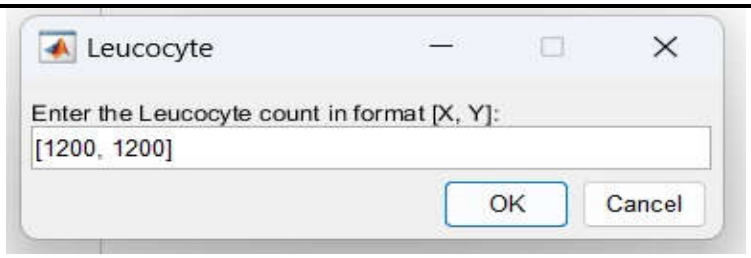
Output predicted class for patient-1 using GMM

Figure 14: Input and Output of Eosinophil count (Patient-1)

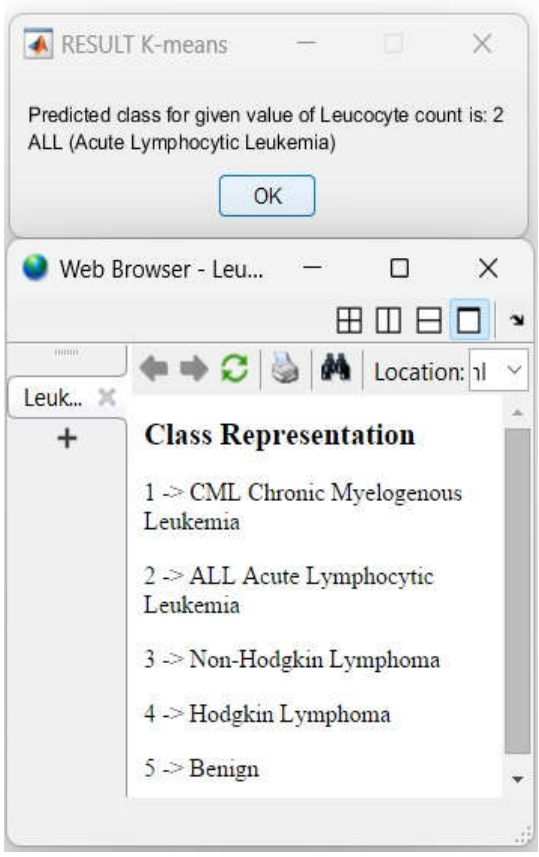
Table 3: Hematological Parameters of Patient-2

(Input for Unsupervised Learning Algorithm)

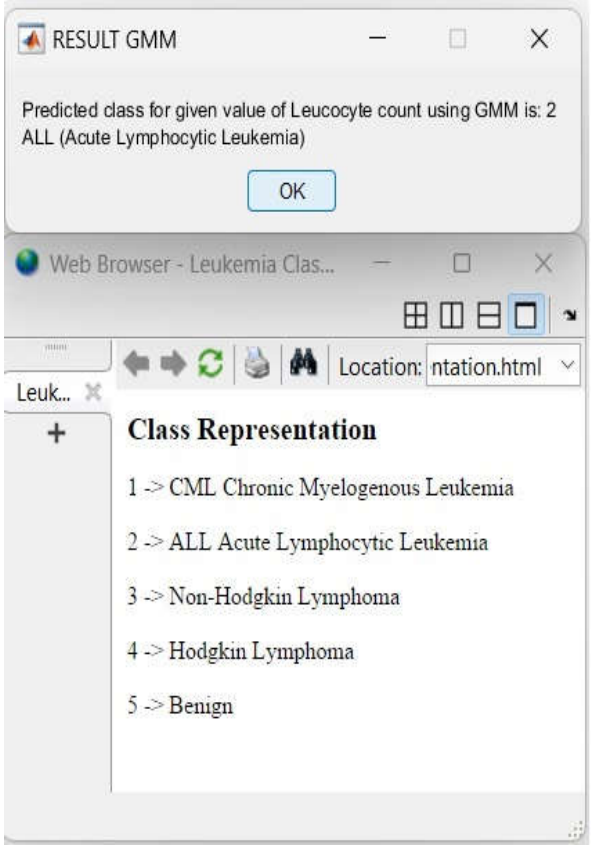
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
1200	14	53	3	7	Acute Lymphocytic Leukemia



Input value of Leucocyte count (Patient-2)

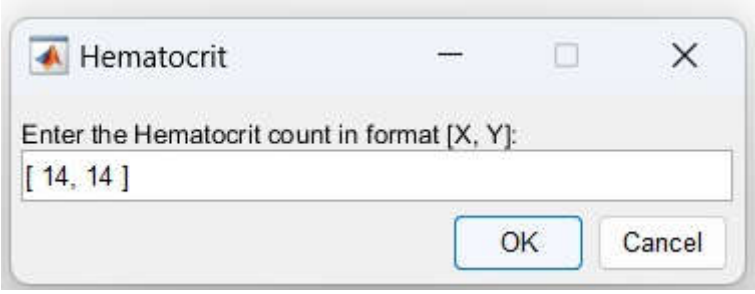


Output predicted class for patient-2 using K-means

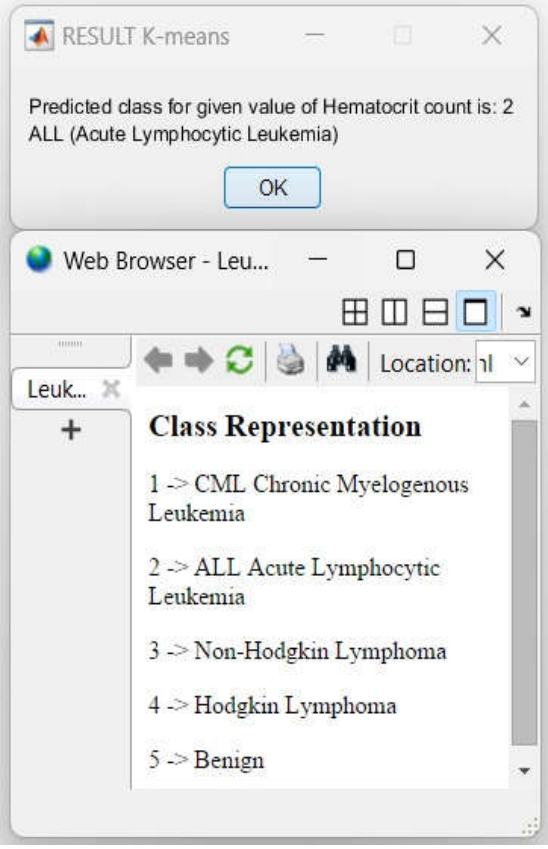


Output predicted class for patient-2 using GMM

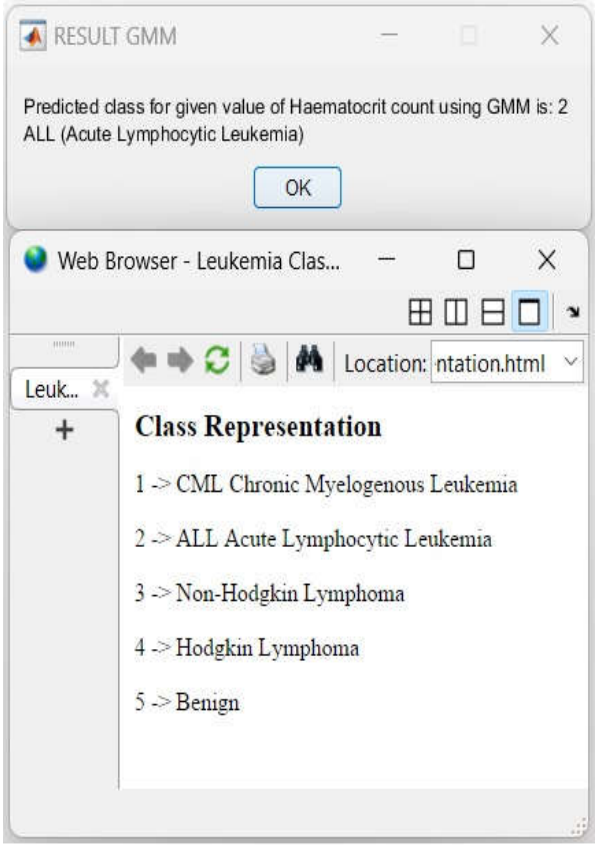
Figure 15: Input and Output of Leucocyte count (Patient-2)



Input value of Hematocrit (Patient-2)

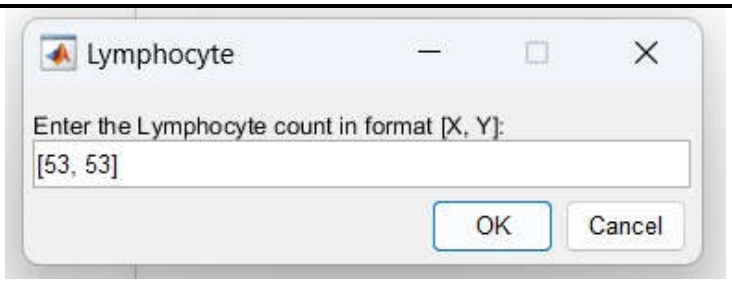


Output predicted class for patient-2 using K-means

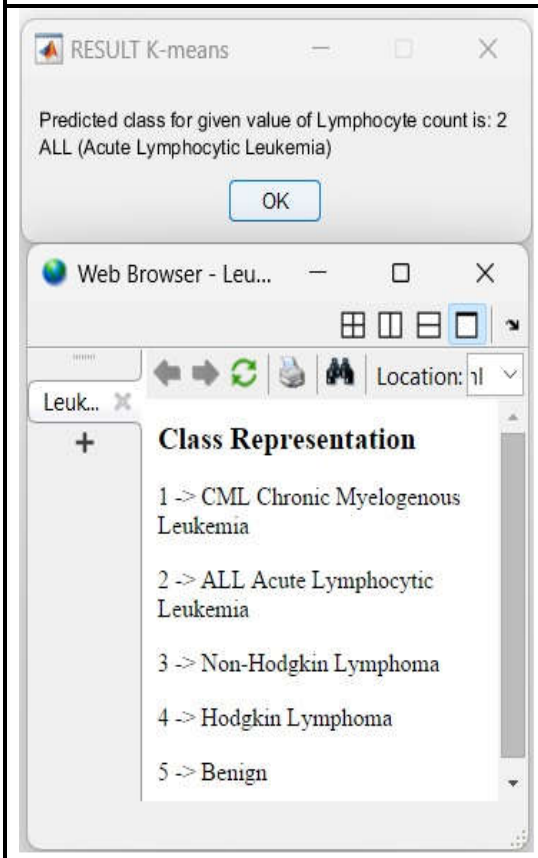


Output predicted class for patient-2 using GMM

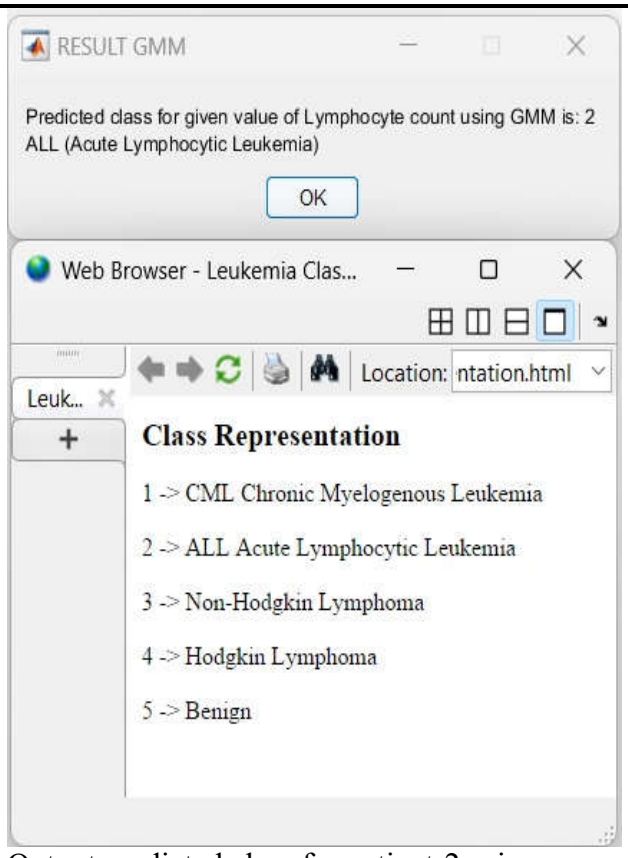
Figure 16: Input and Output of Hematocrit count (Patient-2)



Input value of Lymphocyte Count (Patient-2)

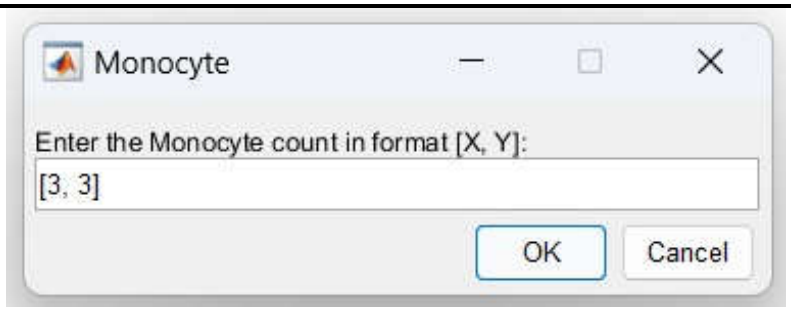


Output predicted class for patient-2 using K-means

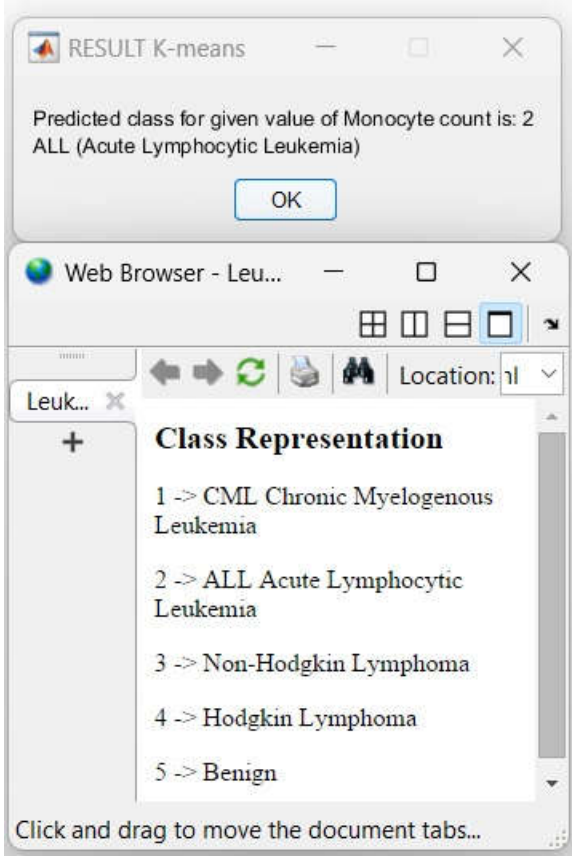


Output predicted class for patient-2 using GMM

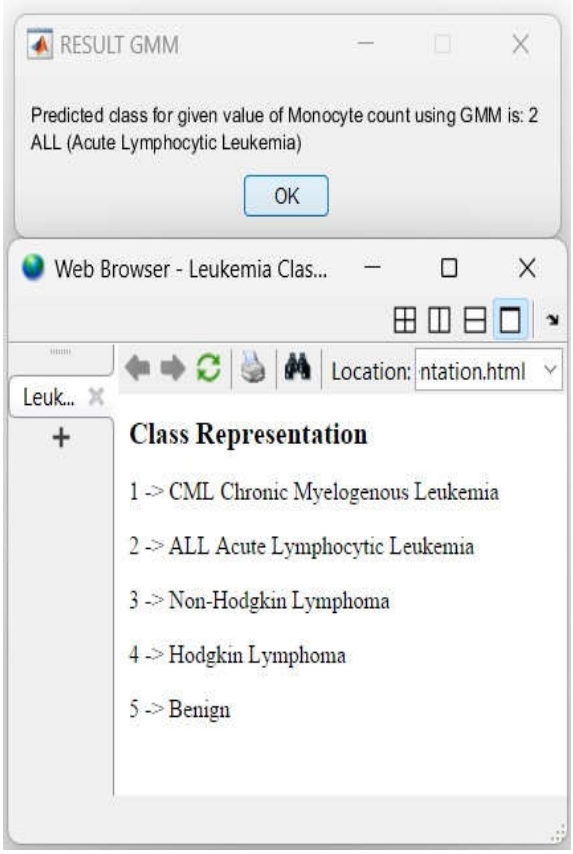
Figure 17: Input and Output of Lymphocyte count (Patient-2)



Input value of Monocyte (Patient-2)

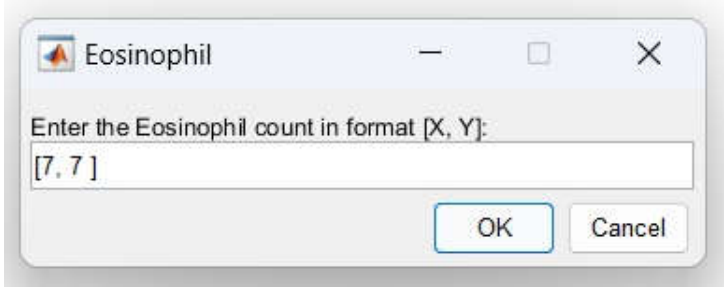


Output predicted class for patient-2 using K-means

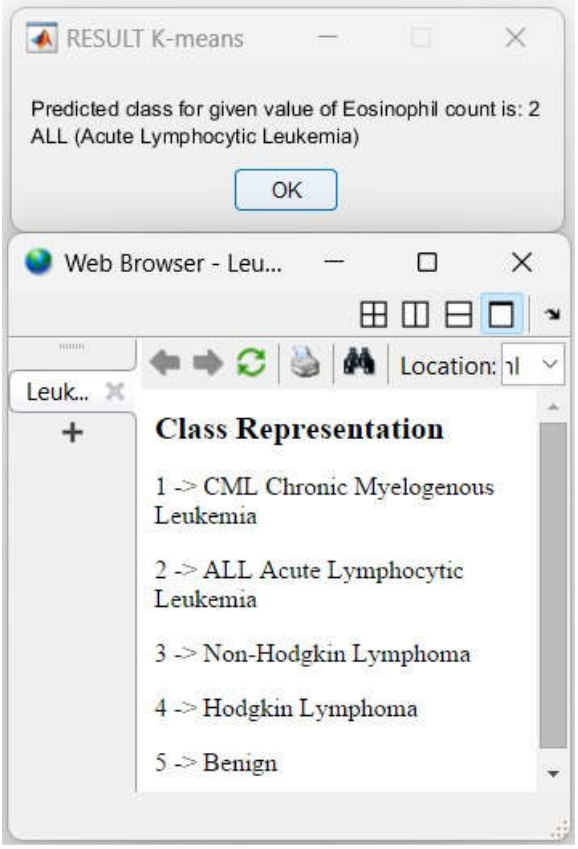


Output predicted class for patient-2 using GMM

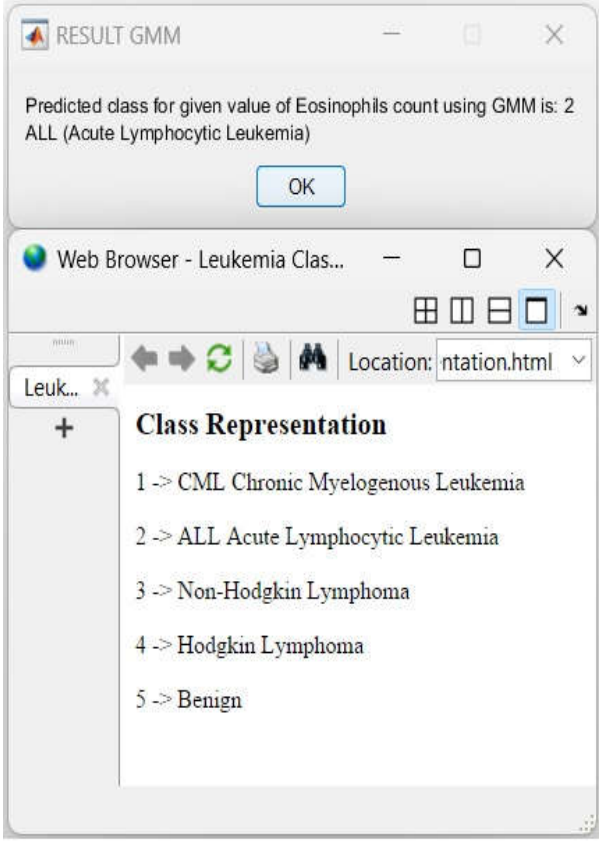
Figure 18: Input and Output of Monocyte count (Patient-2)



Input value of Eosinophil (Patient-2)



Output predicted class for patient-2 using K-means



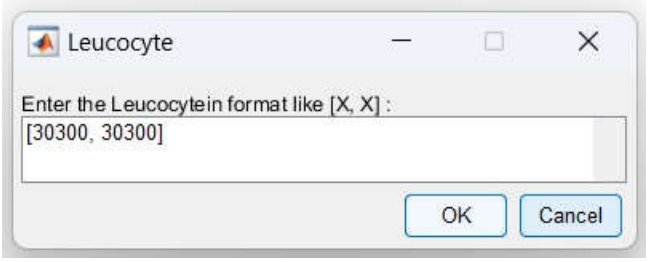
Output predicted class for patient-2 using GMM

Figure 19: Input and Output of Eosinophil count (Patient-2)

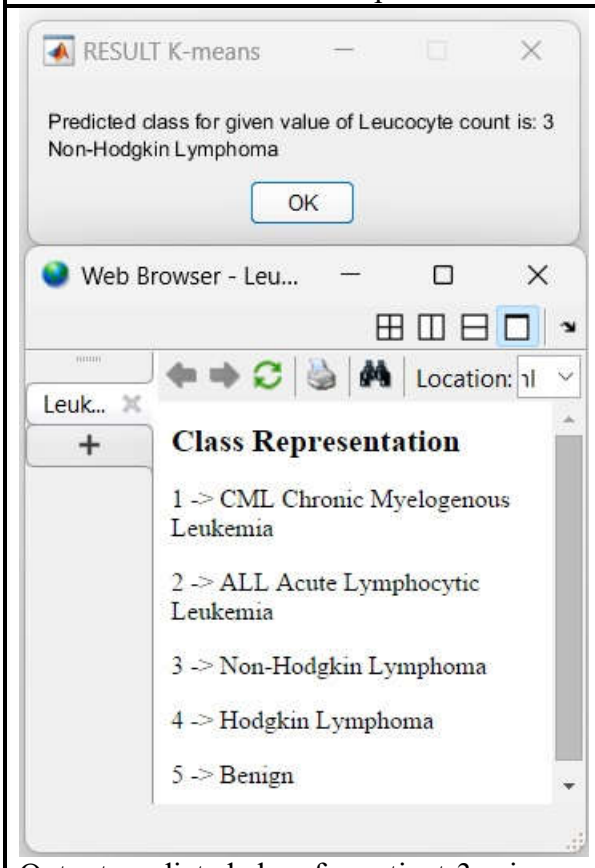
Table 4: Hematological Parameters of Patient-3

(Input for Unsupervised Learning Algorithm)

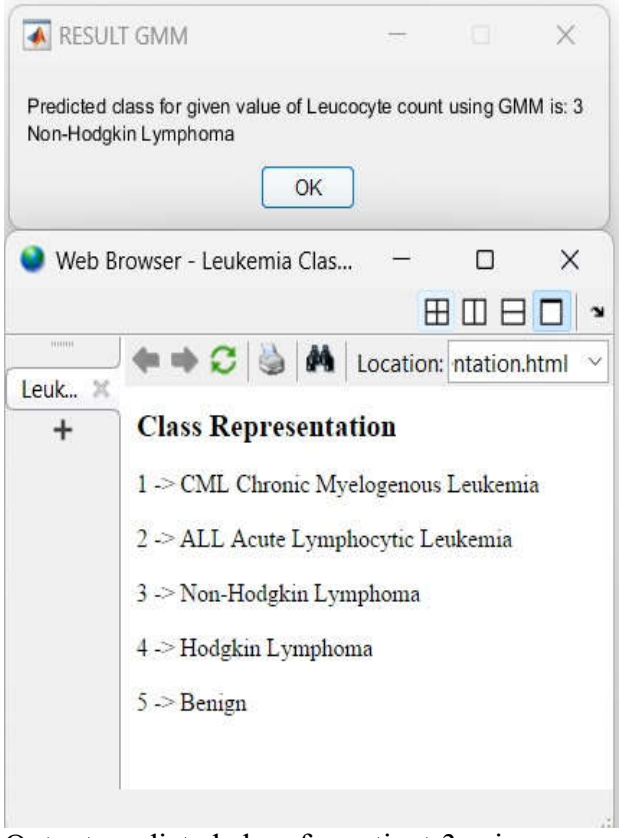
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
30300	34	8	6	11	Non-Hodgkin Lymphoma



Input value of Leucocyte count (Patient-3)

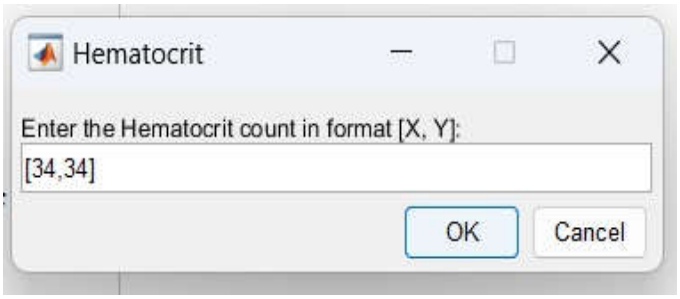


Output predicted class for patient-3 using K-means

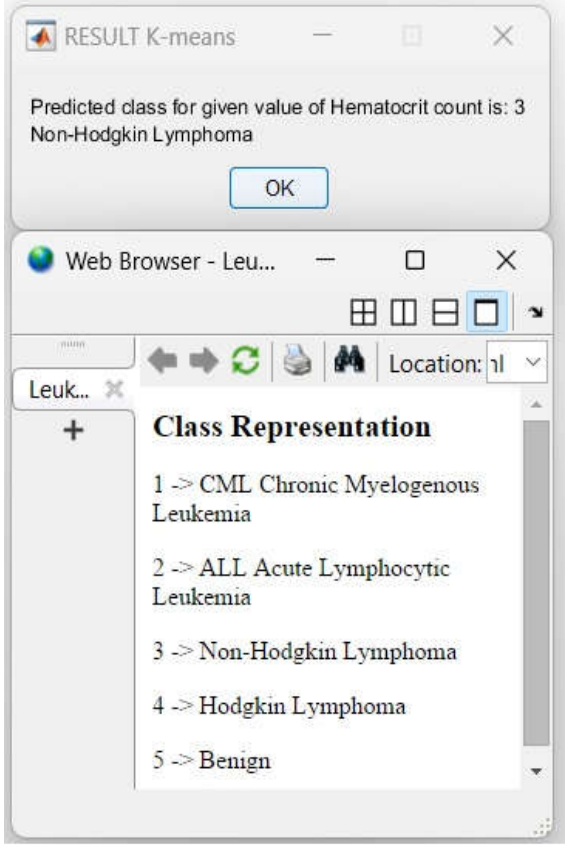


Output predicted class for patient-3 using GMM

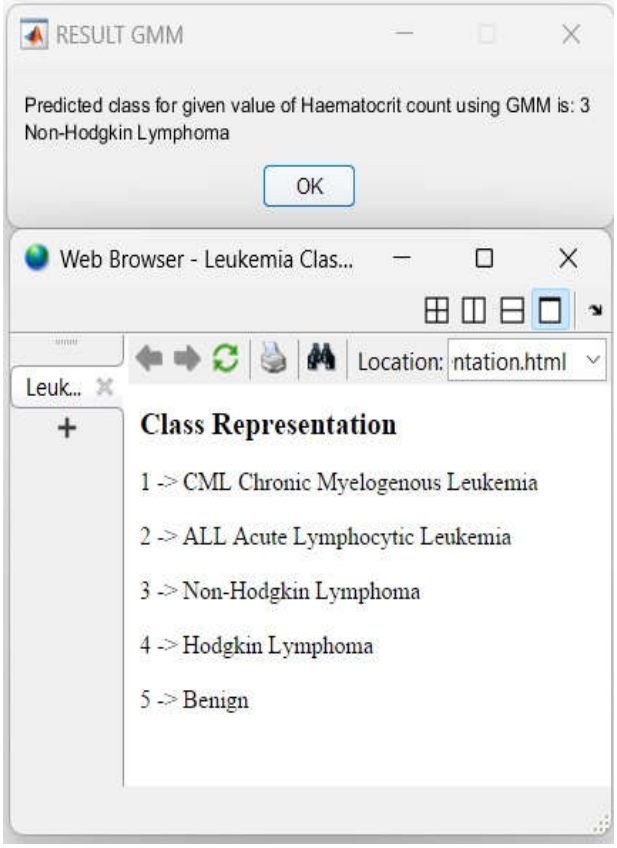
Figure 20: Input and Output of Leucocyte count (Patient-3)



Input value of Hematocrit (Patient-3)

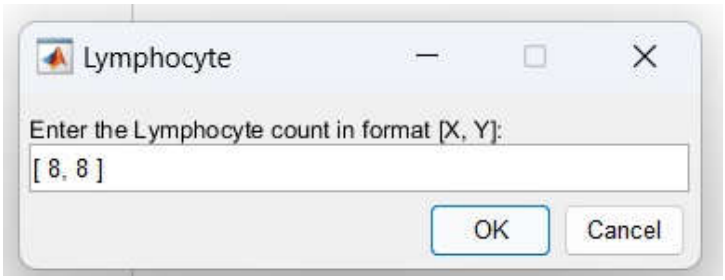


Output predicted class for patient-3 using K-means

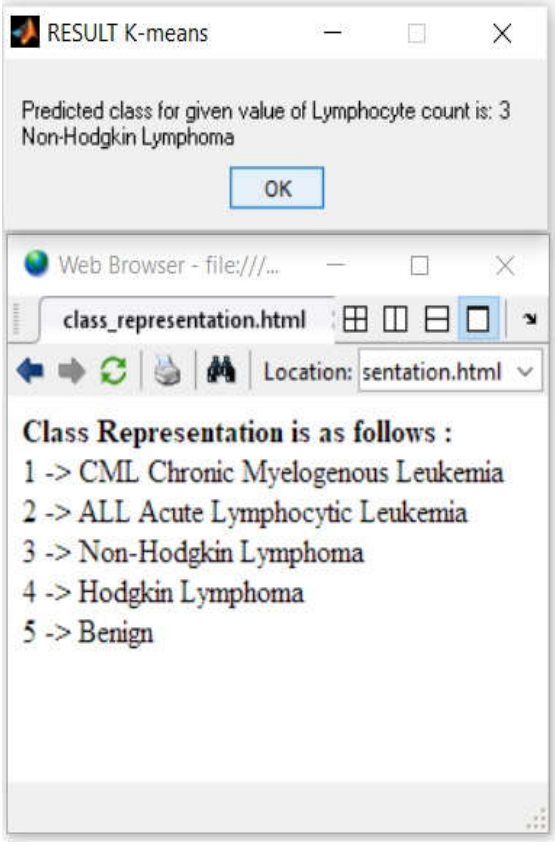


Output predicted class for patient-3 using GMM

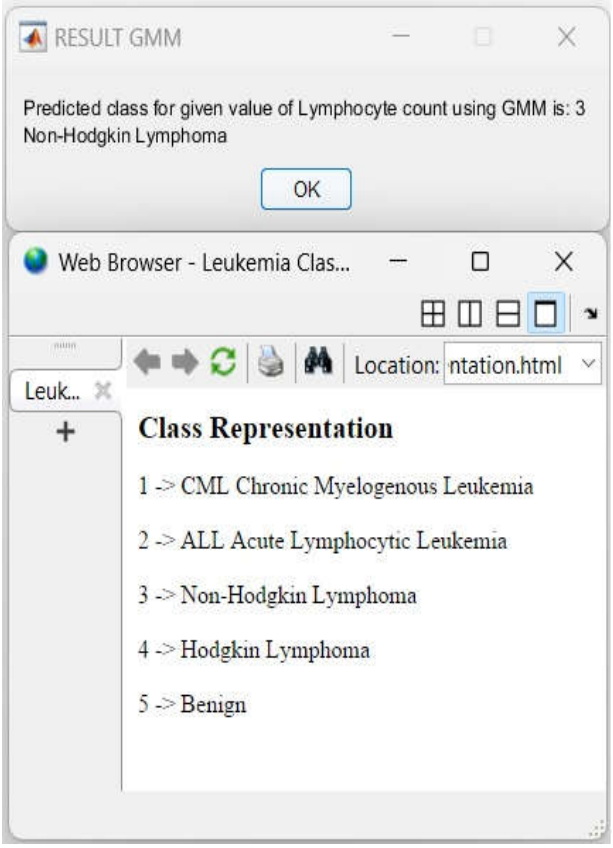
Figure 21: Input and Output of Hematocrit count (Patient-3)



Input value of Lymphocyte Count (Patient-3)

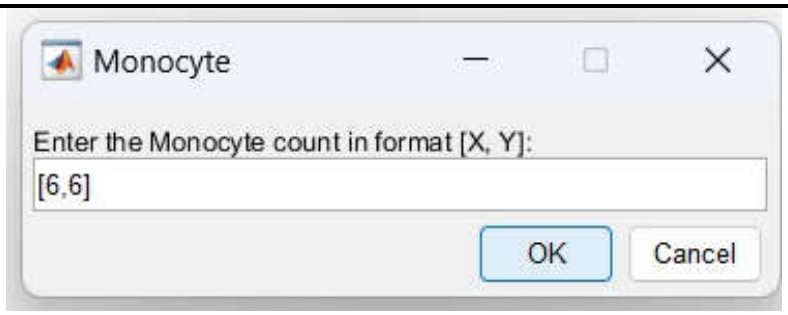


Output predicted class for patient-3 using K-means

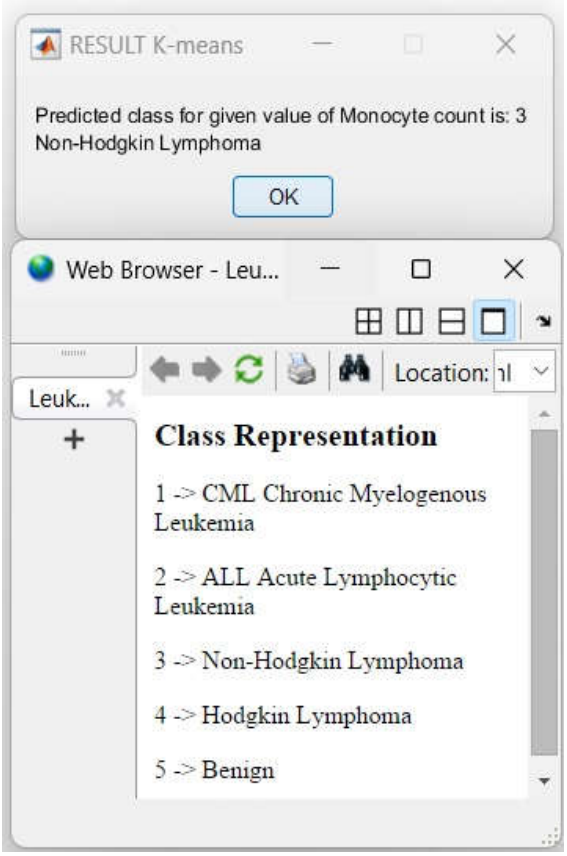


Output predicted class for patient-3 using GMM

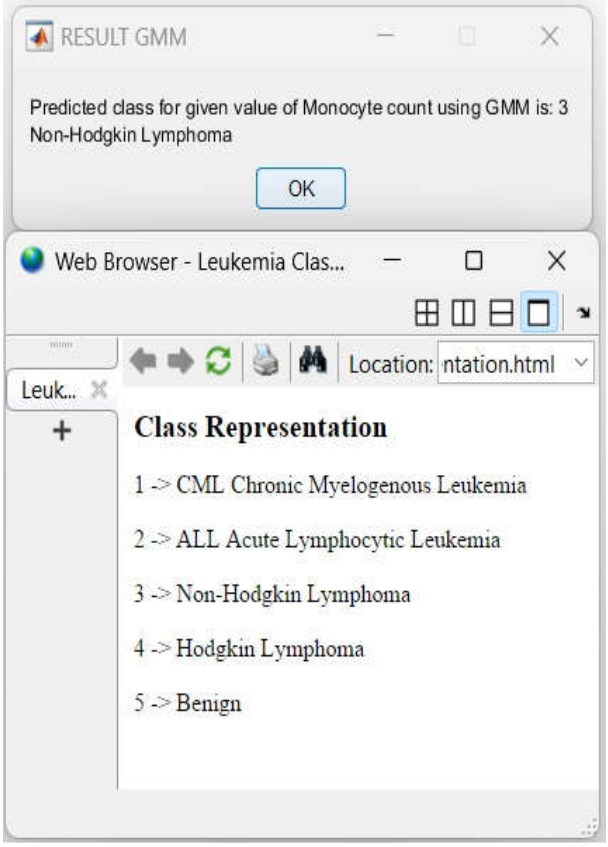
Figure 22: Input and Output of Lymphocyte count (Patient-3)



Input value of Monocyte (Patient-3)

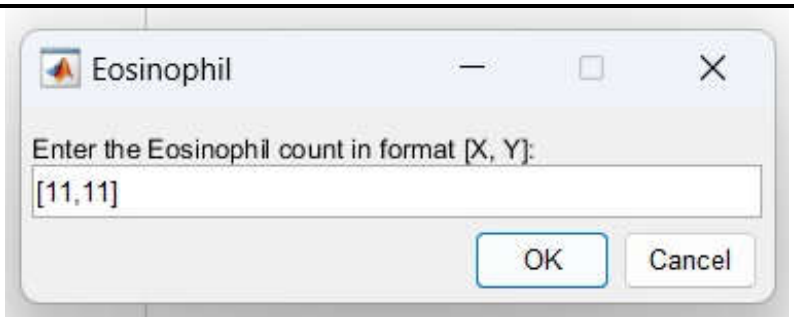


Output predicted class for patient-3 using K-means

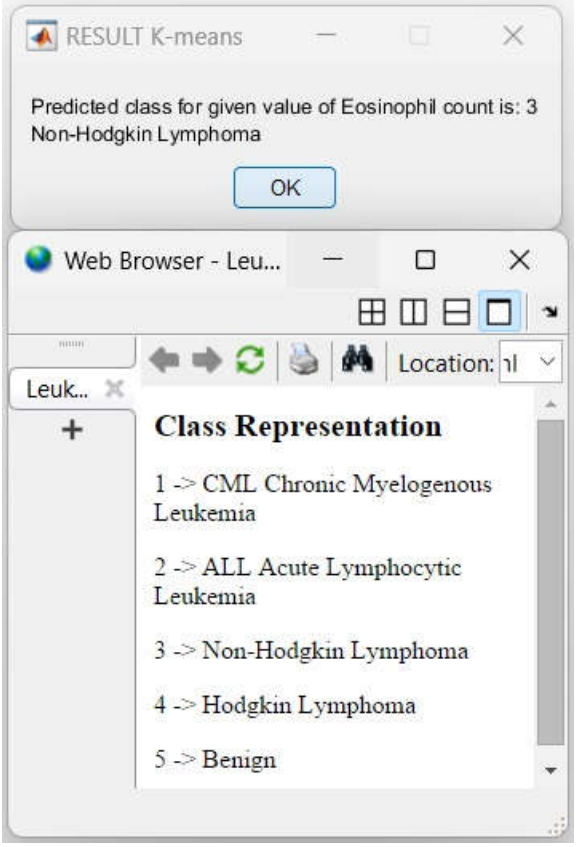


Output predicted class for patient-3 using GMM

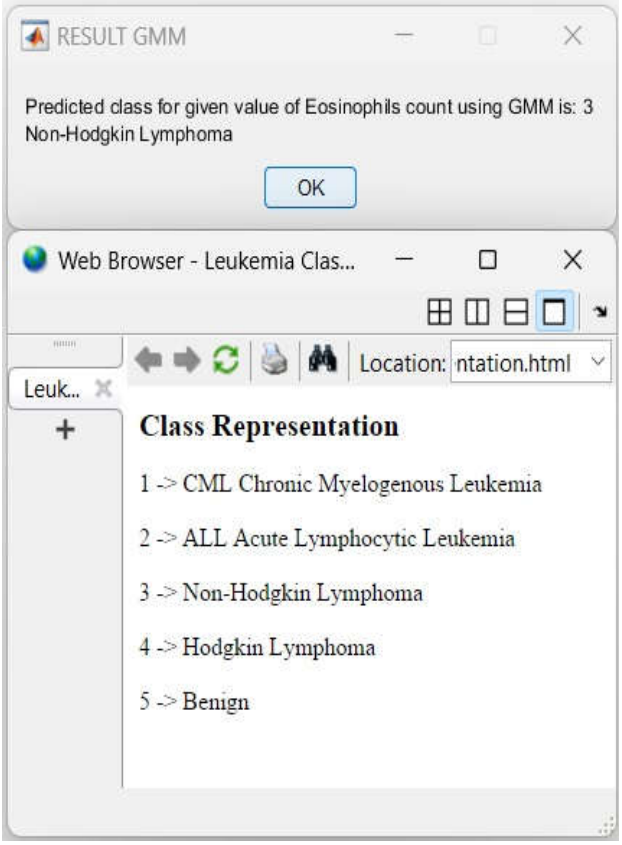
Figure 23: Input and Output of Monocyte count (Patient-3)



Input value of Eosinophil (Patient-3)



Output predicted class for patient-3 using K-means

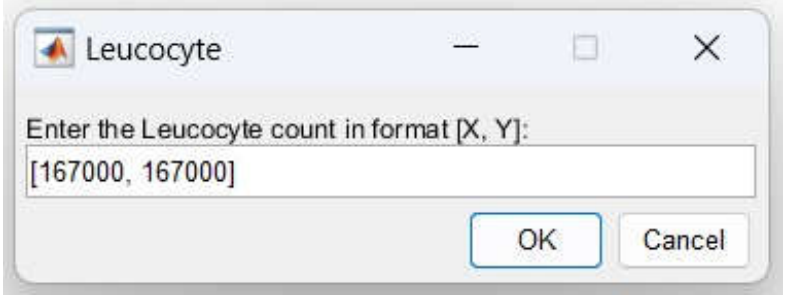


Output predicted class for patient-3 using GMM

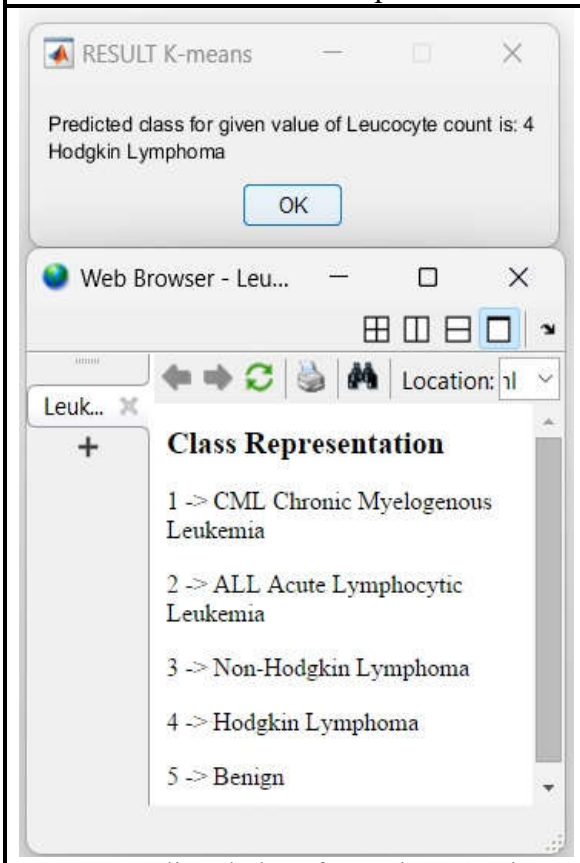
Figure 24: Input and Output of Eosinophil count (Patient-3)

Table 5: Hematological Parameters of Patient-4 (Input for Unsupervised Learning Algorithm)

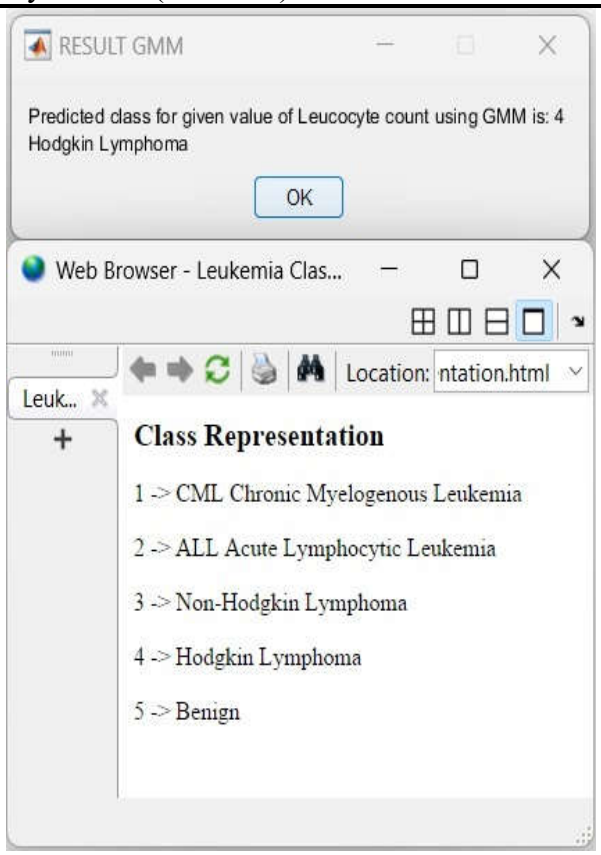
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
167000	19	70	22	1	Hodgkin Lymphoma



Input value of Leucocyte count (Patient-4)

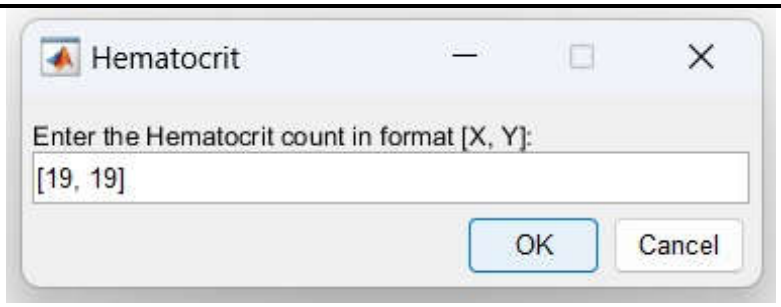


Output predicted class for patient-4 using K-means

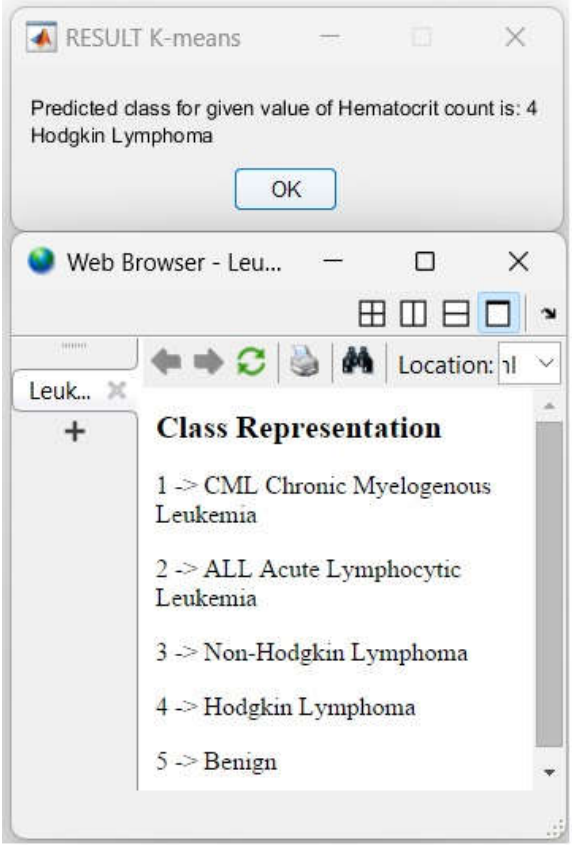


Output predicted class for patient-4 GMM

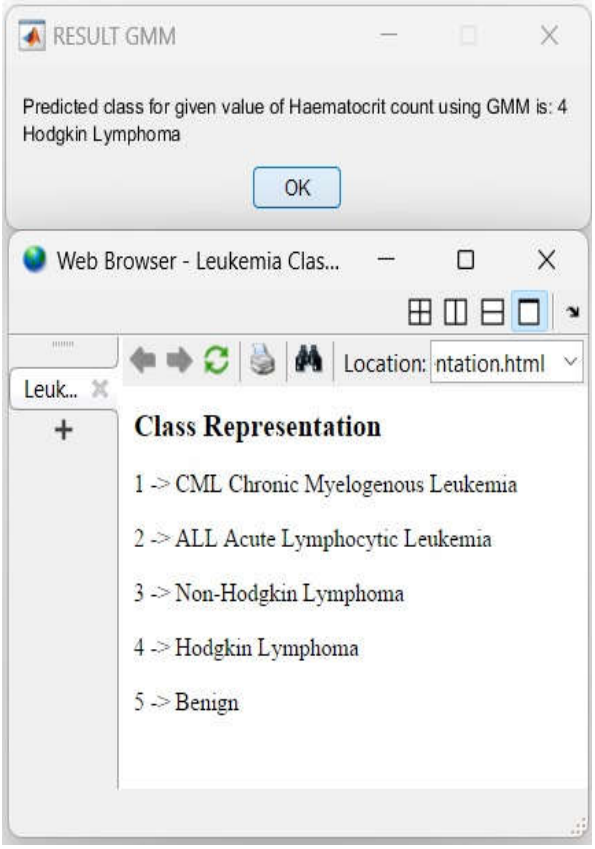
Figure 25: Input and Output of Leucocyte count (Patient-4)



Input value of Hematocrit (Patient-4)

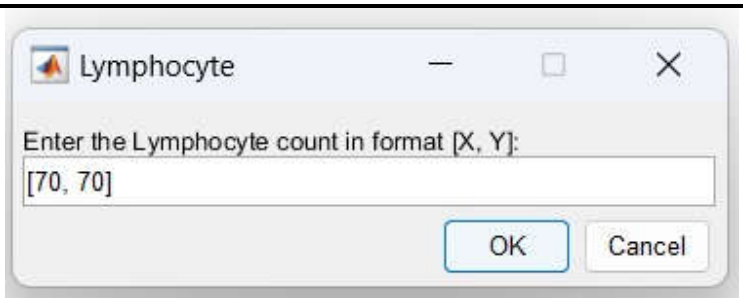


Output predicted class for patient-4 using K-means

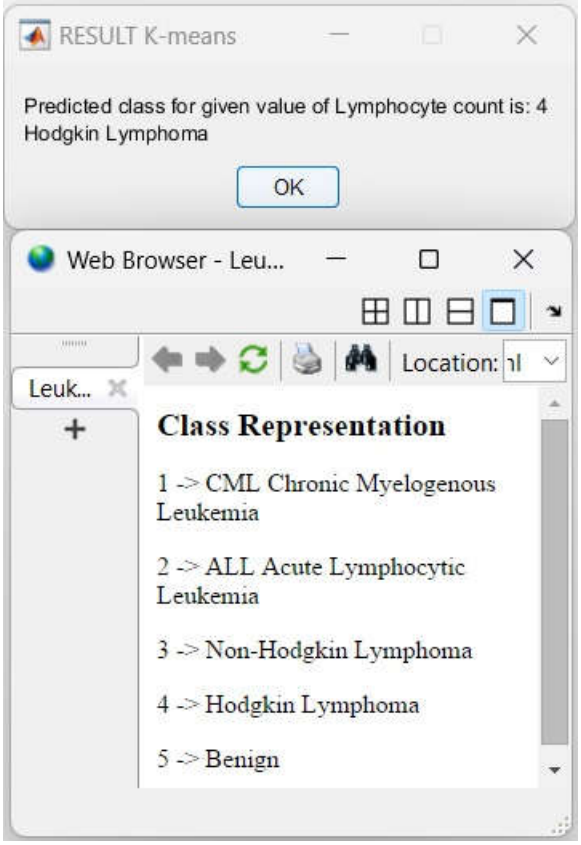


Output predicted class for patient-4 GMM

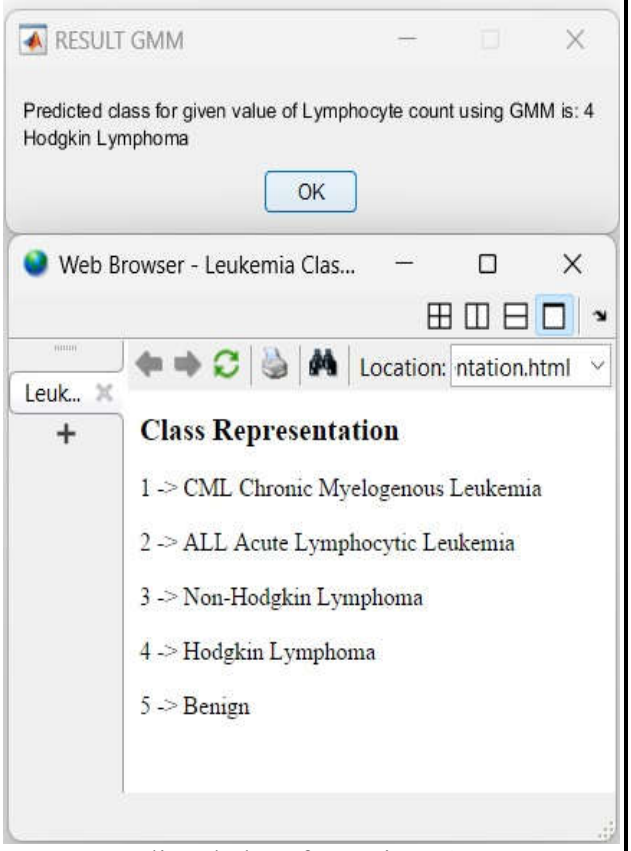
Figure 26: Input and Output of Hematocrit count (Patient-4)



Input value of Lymphocyte Count (Patient-4)

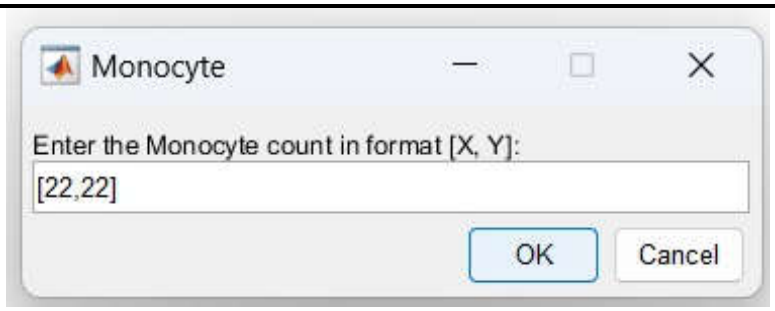


Output predicted class for patient-4 using K-means

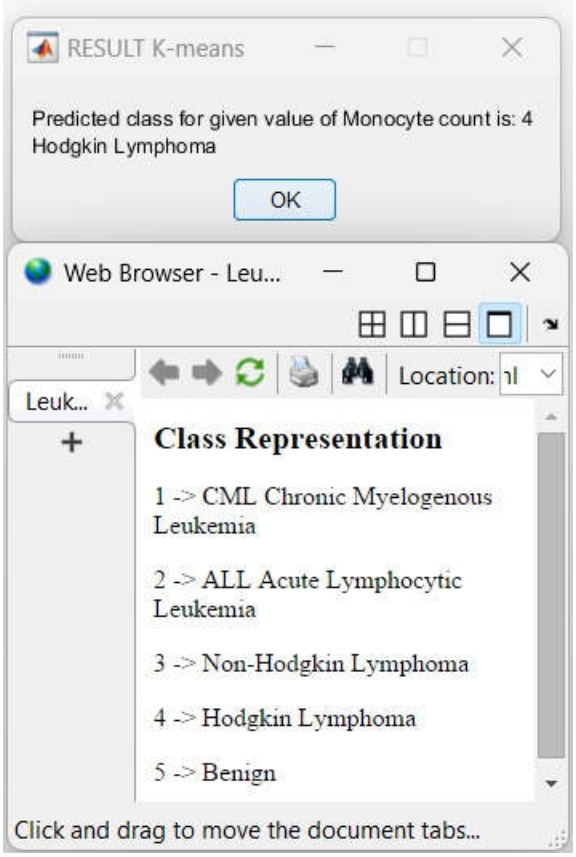


Output predicted class for patient-4 GMM

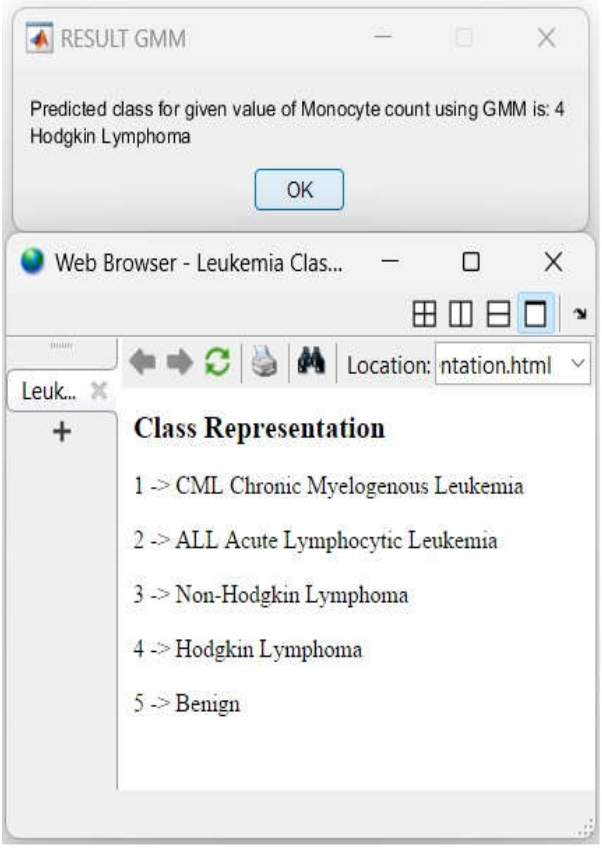
Figure 27: Input and Output of Lymphocyte count (Patient-4)



Input value of Monocyte (Patient-4)

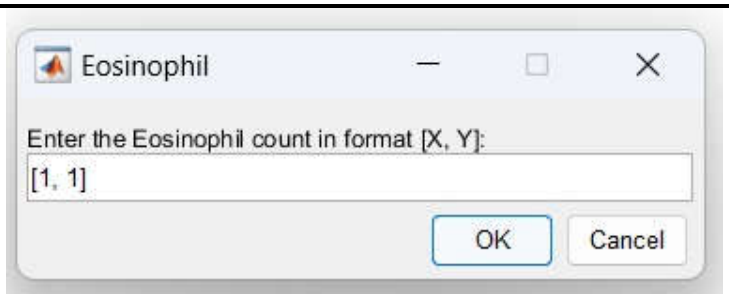


Output predicted class for patient-4 using K-means

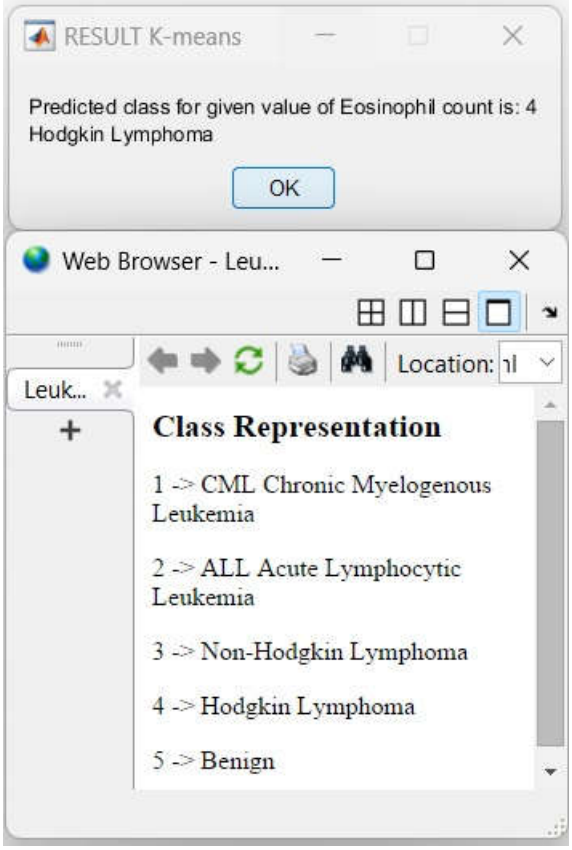


Output predicted class for patient-4 GMM

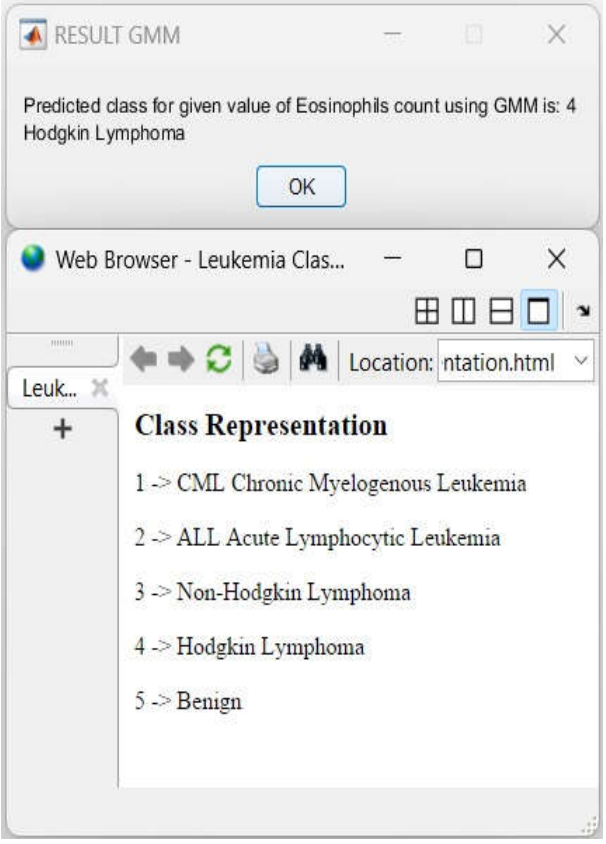
Figure 28: Input and Output of Monocyte count (Patient-4)



Input value of Eosinophil (Patient-4)



Output predicted class for patient-4 using K-means



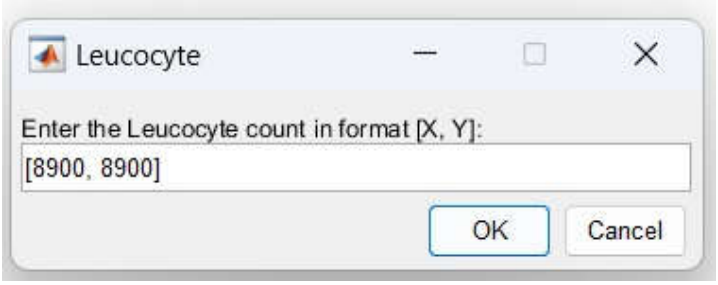
Output predicted class for patient-4 GMM

Figure 29: Input and Output of Eosinophil count (Patient-4)

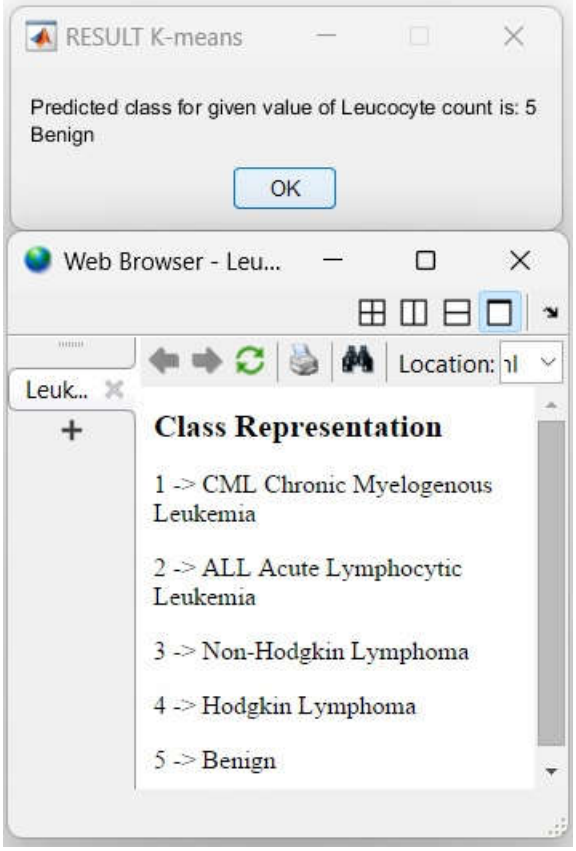
Table 6: Hematological Parameters of Patient-5

(Input for Unsupervised Learning Algorithm)

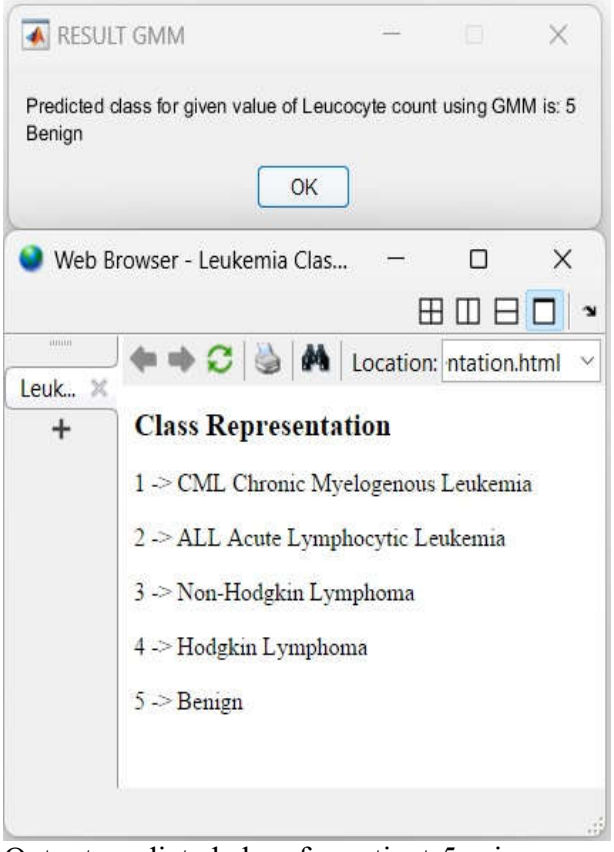
Total Leucocyte count	Haematocrit	Lymphocyte	Monocyte	Eosinophils	Class/ (Diagnosed with Type)
8900	44	36	5	2	Benign



Input value of Leucocyte count (Patient-5)

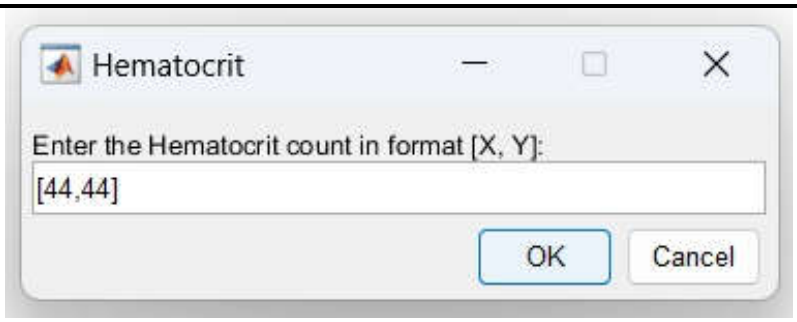


Output predicted class for patient-5 using K-means

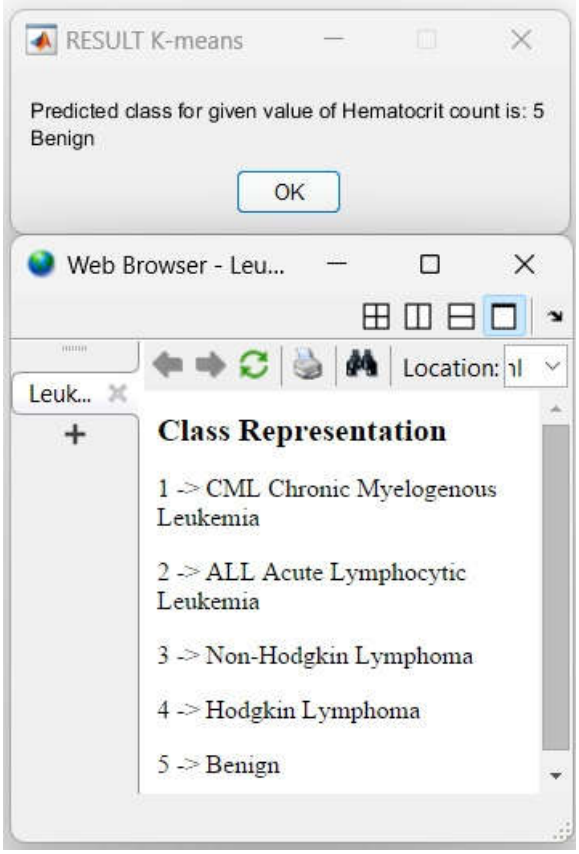


Output predicted class for patient-5 using GMM

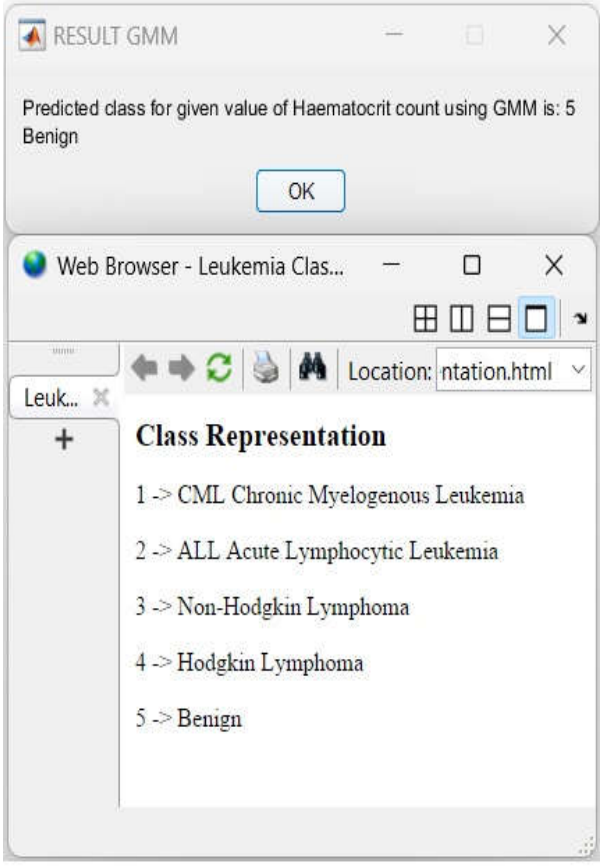
Figure 30: Input and Output of Leucocyte count (Patient-5)



Input value of Hematocrit (Patient-5)

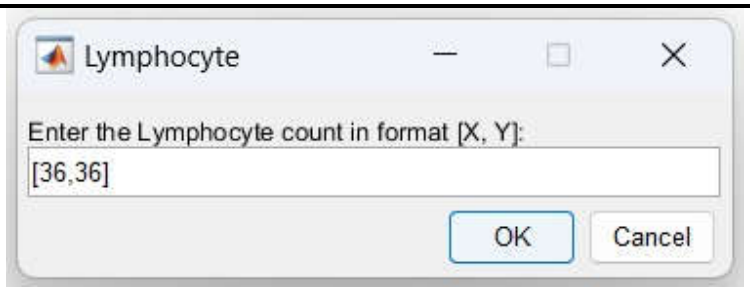


Output predicted class for patient-5 using K-means

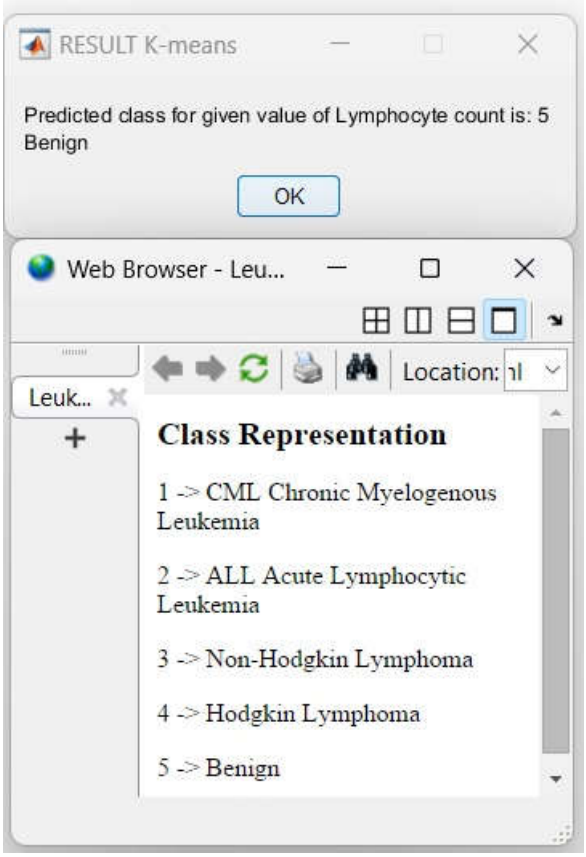


Output predicted class for patient-5 using GMM

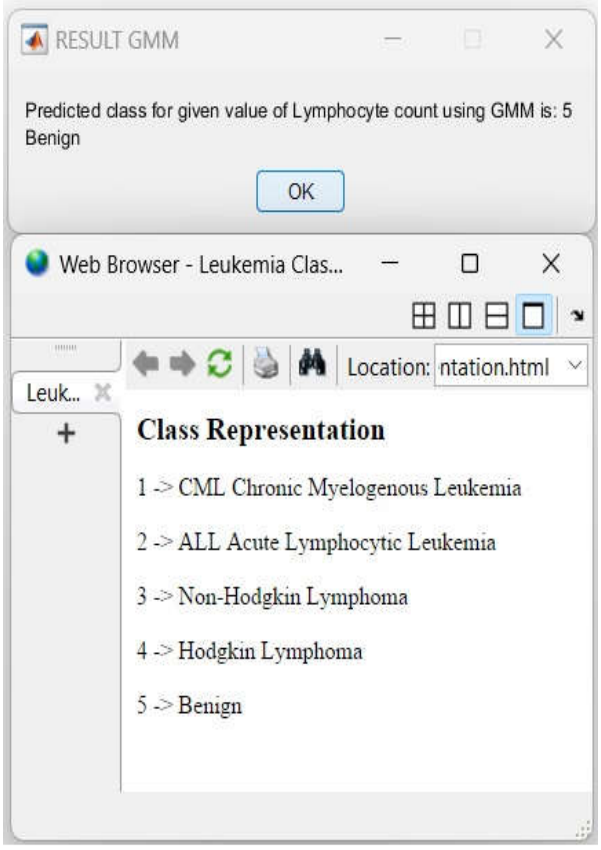
Figure 31: Input and Output of Hematocrit count (Patient-5)



Input value of Lymphocyte Count (Patient-5)

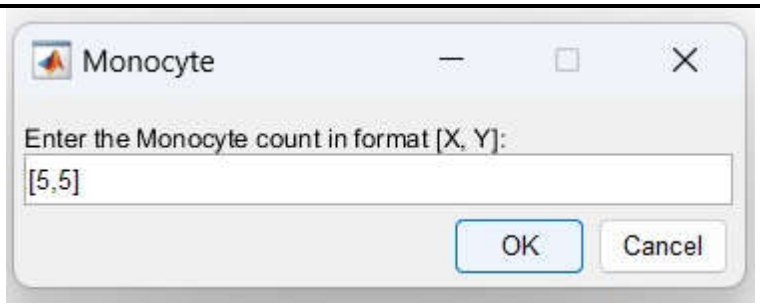


Output predicted class for patient-5 using K-means

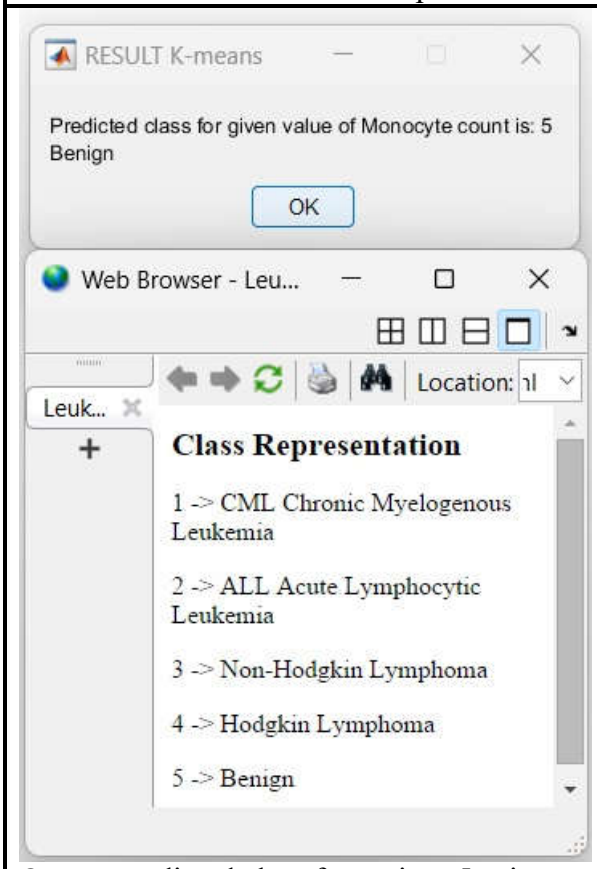


Output predicted class for patient-5 using GMM

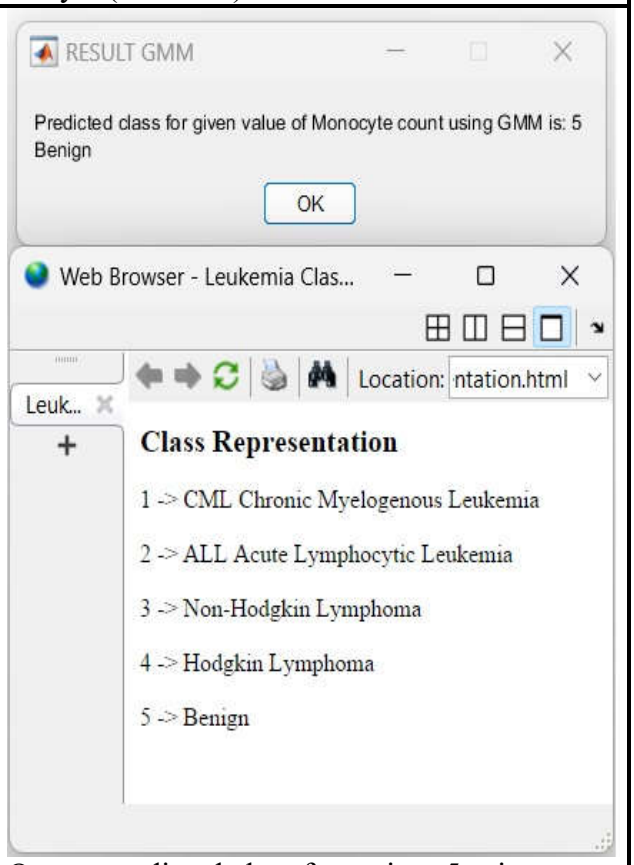
Figure 32: Input and Output of Lymphocyte count (Patient-5)



Input value of Monocyte (Patient-5)

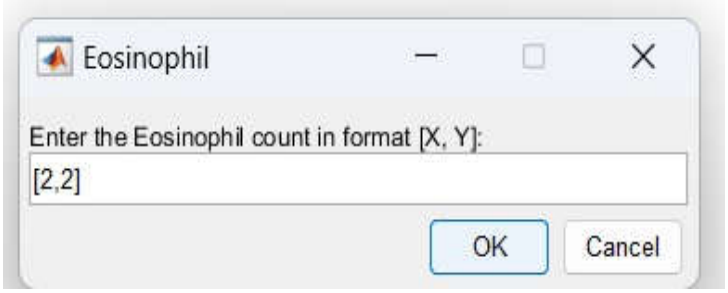


Output predicted class for patient-5 using K-means

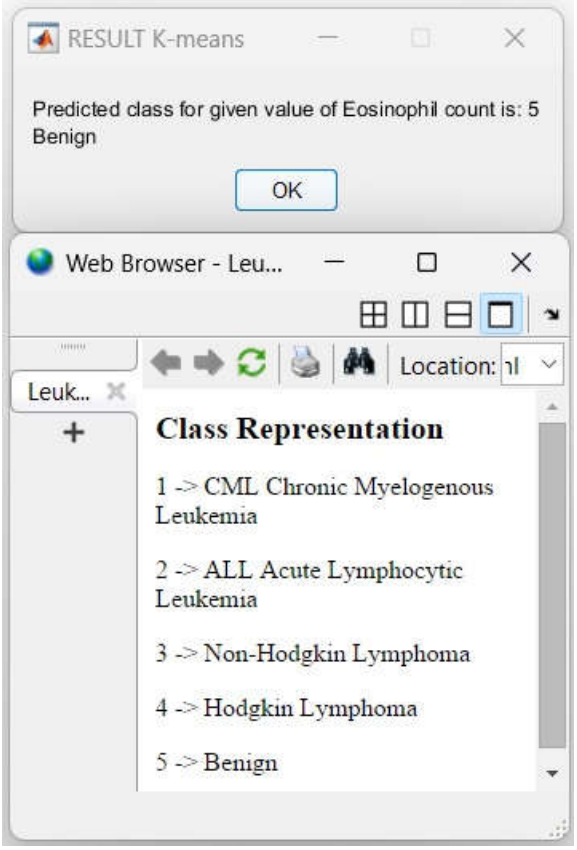


Output predicted class for patient-5 using GMM

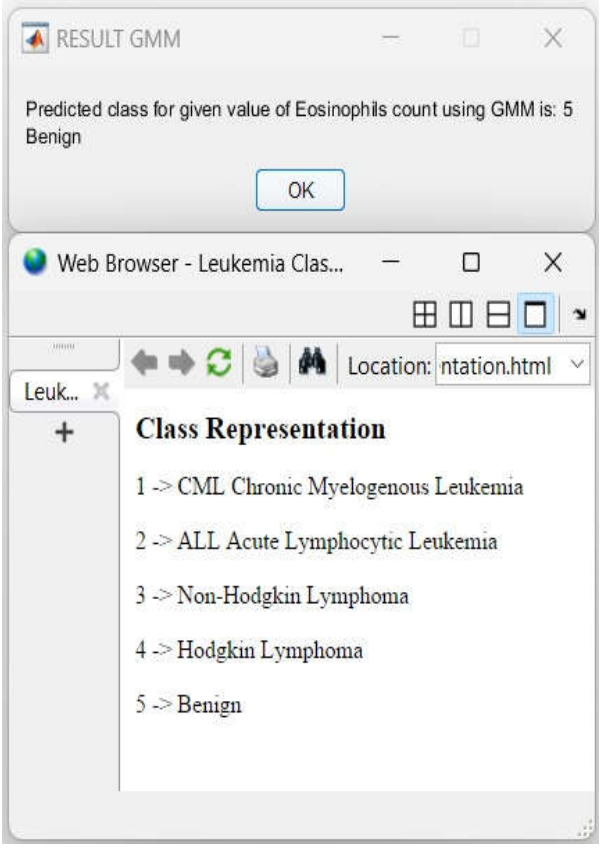
Figure 33: Input and Output of Monocyte count (Patient-5)



Input value of Eosinophil (Patient-5)



Output predicted class for patient-5 using K-means



Output predicted class for patient-5 using GMM

Figure 34: Input and Output of Eosinophil count (Patient-5)

The predicted results are highly promising for both algorithms, as the assigned classifications align with the diagnoses made by medical experts, accurately identifying patients with the same type of disease.

Comparative Analysis of K-means and GMM Algorithms

We compare both the algorithms by using three performance metrics such as accuracy, precision and recall. Accuracy measures the proportion of correctly classified data points compared to the total number of data points.

Mathematically, accuracy is defined as:

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

The Accuracy versus number of neighbours (k) graph is shown in the figure 35 below. The value of accuracy for the values of neighbours 3-5 k is around 85% which is quite satisfying.

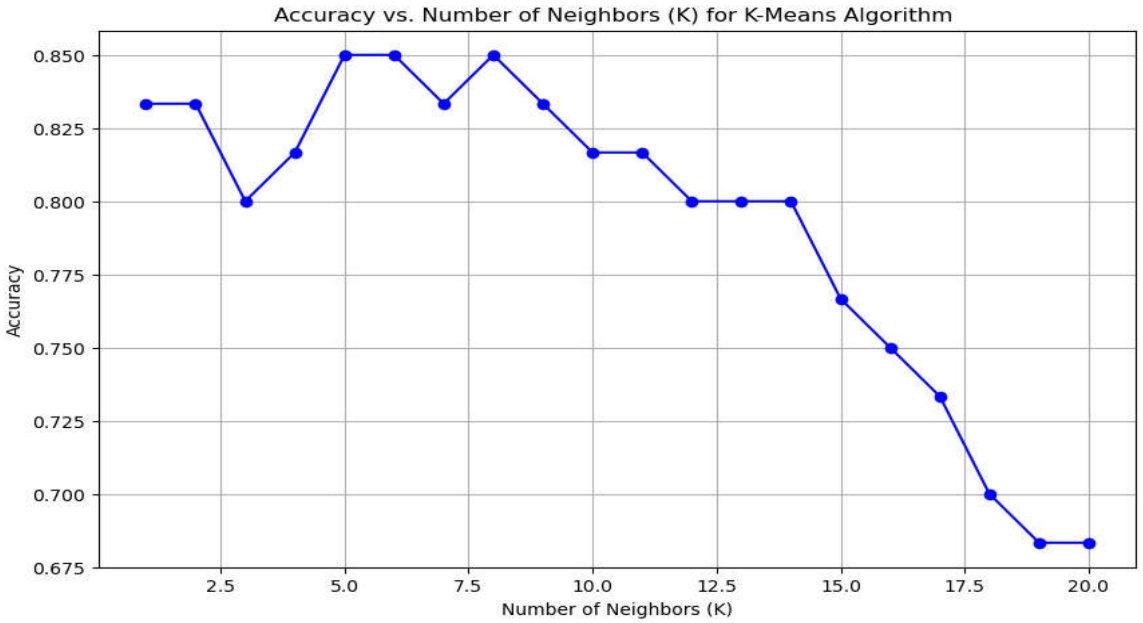


Figure 35: Accuracy performance of K- Means Algorithm

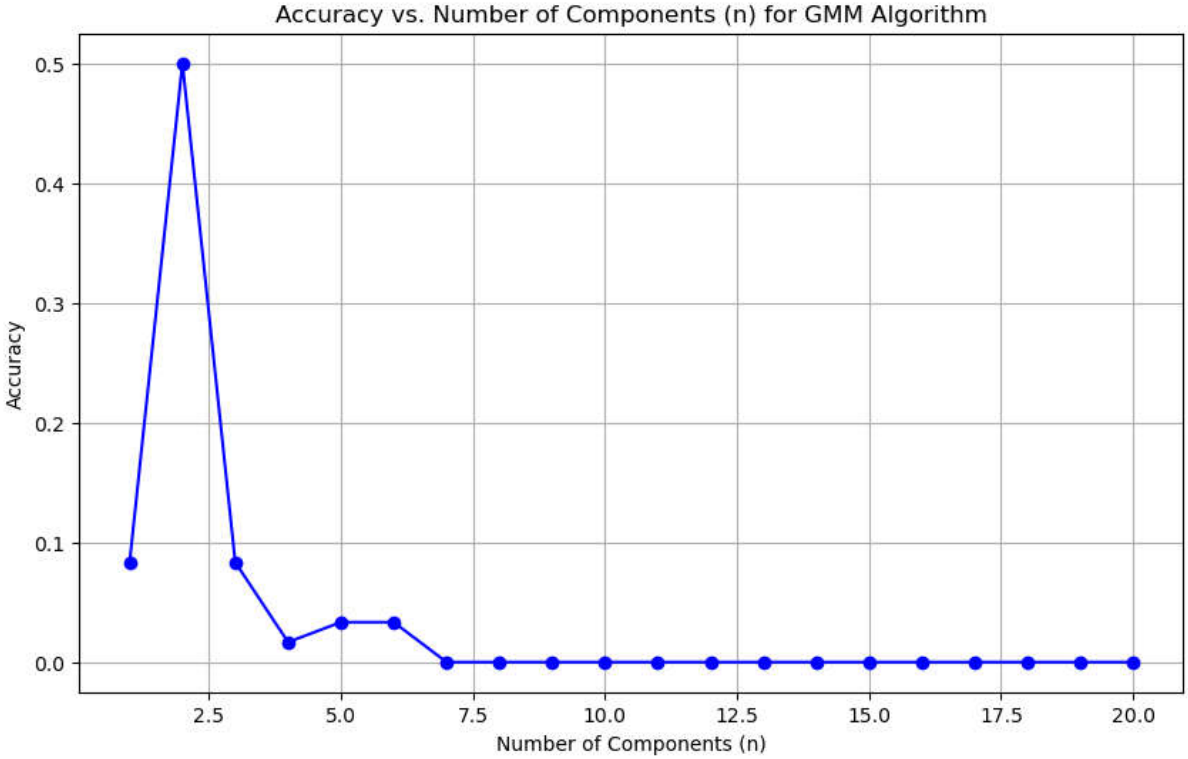


Figure 36: Accuracy performance of GMM Algorithm

The value of accuracy for GMM reaches around 50% which is quite less than K-means.

Precision and recall are evaluation metrics commonly used in the context of classification algorithms to assess their performance. Precision is a measure that indicates the accuracy of the positive predictions made by a classifier. It represents the ratio of correctly predicted positive observations to the total predicted positive observations. In other words, it assesses the model's ability to not label a negative sample as positive (Sinaga et al. 2023).

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 37 the precision is about 83% and recall is 85%, which is quite satisfying.

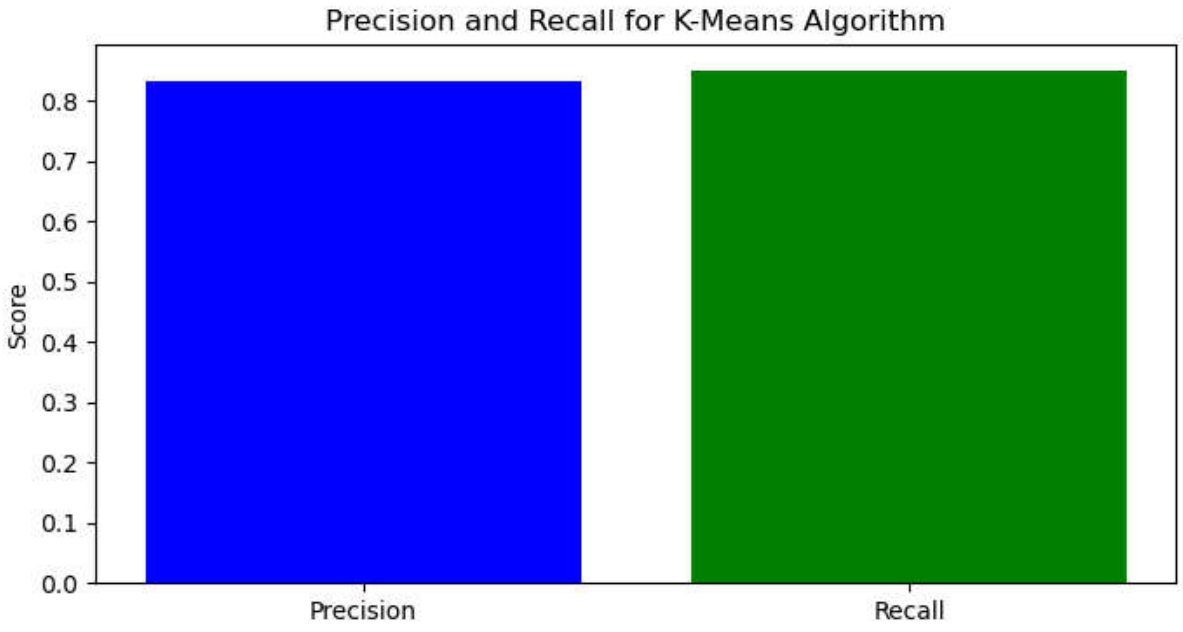


Figure 37: Precision and Recall performance of K- Means Algorithm

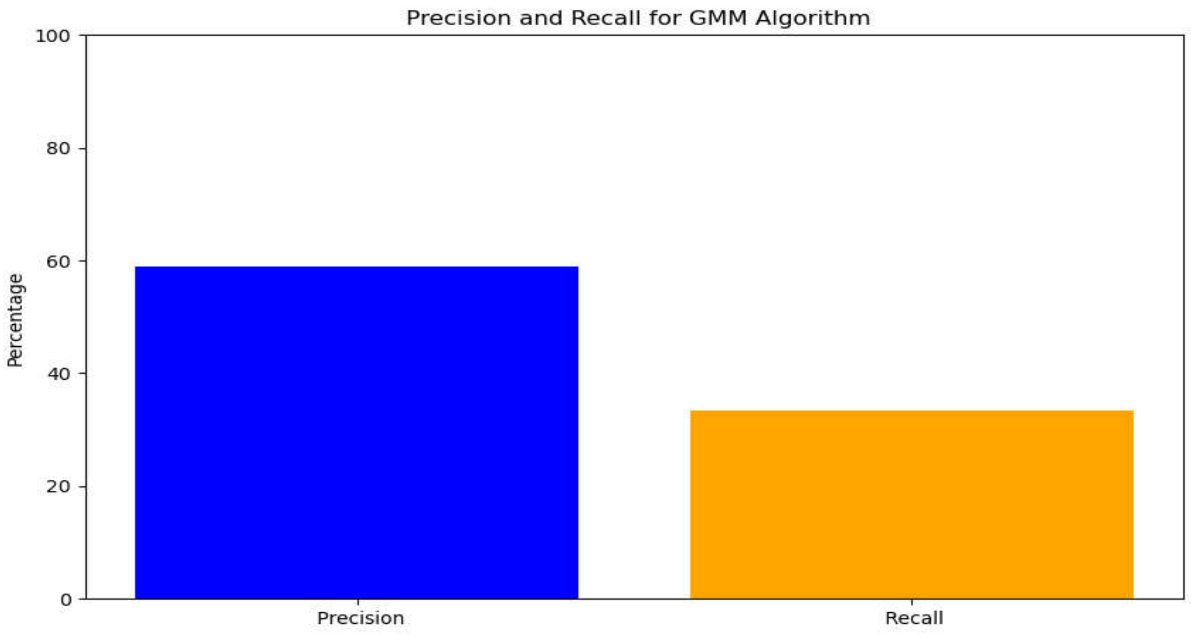


Figure 38: Precision and Recall performance for GMM

The Recall value of GMM is 33% which is quite less than K-means which has 85%.

Table 7: Performance Parameter Comparison Table

Algorithm	Accuracy	Precision	Recall
K-means	85.00%	83.31%	85.00%
GMM	50.1%	59.0%	33.3 %

K-means outperformed GMM in leukemia diagnosis because it effectively minimizes intra-cluster variance, leading to more distinct cluster formations. GMM, being a probabilistic model, can struggle with overlapping clusters, affecting classification performance. The experimental results showed that K-means achieved higher accuracy, precision, and recall, indicating its robustness in segmenting the dataset for leukemia classification. Additionally, K-means is computationally efficient, ensuring faster convergence compared to the iterative expectation-maximization process of GMM.

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 four classes of cancer along with benign cases on blood parameters, with the K-means algorithm demonstrating effective performance as evaluated through metrics including accuracy, precision, and recall.

References

Abid, Aymen, Adel Thajaoui, Zeadally Sherali, Moahmed Miladi, and Abdennaceur Kachouri (2023), Smart K Nearest Neighbor Outlier Detection for Electroencephalogram Signal. https://doi.org/10.21203/rs.3.rs-3005782/v1.

Adinanta, Hendra, Edi Kurniawan, and Jalu A. Prakosa (2020), Physical Distancing Monitoring with Background Subtraction Methods. In *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, 45–50. IEEE. https://doi.org/10.1109/ICRAMET51080.2020.9298687.

Archana, K. N., and V. Kumar (2023), GMM-Based Mean Approximation for Skin Disease Classification Using Machine Learning. In *2023 International Conference on Computational Intelligence, Networks and Security (ICCINS)*, 1–5. IEEE. https://doi.org/10.1109/ICCINS58907.2023.10450027.

Blackadar, C. B. (2016), Historical Review of the Causes of Cancer. *World Journal of Clinical Oncology* 7(1): 54–86. https://doi.org/10.5306/wjco.v7.i1.54.

Chen, Yongliang, and Alina Shayilan (2022), Dictionary Learning for Multivariate Geochemical Anomaly Detection for Mineral Exploration Targeting. *Journal of Geochemical Exploration* 235: 106958. https://doi.org/10.1016/j.gexplo.2022.106958.

Haibin, Zhang, Xiao Han, Yi Cancan, and Yuan Rui (2023), A Novel Borderline Over-Sampling Method Based on KNN and Deep Gaussian Mixture Model for Imbalanced Data. *Data Analysis and Knowledge Discovery* 7(5): 116–122. https://doi.org/10.11925/infotech.2096-3467.2022.0609.

Hemachandira, V. S., and R. Viswanathan (2022), A Framework on Performance Analysis of Mathematical Model-Based Classifiers in Detection of Epileptic Seizure from EEG Signals

with Efficient Feature Selection. *Journal of Healthcare Engineering* 2022. https://doi.org/10.1155/2022/7654666.

Kalaiyarasi, M., and Harikumar Rajaguru (2022), Performance Analysis of Ovarian Cancer Detection and Classification for Microarray Gene Data. *BioMed Research International* 2022. https://doi.org/10.1155/2022/6750457.

Kalaiyarasi, M., Harikumar Rajaguru, and Saravanan Ravi (2023), Respiratory Disorder Detection and Performance Analysis from PPG Signals Using Short-Time Fourier Transform (STFT) Approach. In *2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR)*, vol. 1, 1–6. IEEE. https://doi.org/10.1109/STCR59085.2023.10396883.

Lin, Huang, Hakimeh Purmehdi, Xiaoning Fei, Yuxin Zhao, Alka Isac, Habib Louafi, and Wei Peng (2023), Two-Stage Clustering for Improve Indoor Positioning Accuracy. *Automation in Construction* 154: 104981. https://doi.org/10.1016/j.autcon.2023.104981.

Pandya, Vandana, Urvashi P. Shukla, and Amit M. Joshi (2023), Novel Features Extraction from EEG Signals for Epilepsy Detection Using Machine Learning Model. *IEEE Sensors Letters* 2023. https://doi.org/10.1109/LSENS.2023.3309254.

Rasul, Rownak Ara, Promy Saha, Diponkor Bala, S. M. Rakib Ul Karim, Md Ibrahim Abdullah, and Bishwajit Saha (2024), An Evaluation of Machine Learning Approaches for Early Diagnosis of Autism Spectrum Disorder. *Healthcare Analytics* 5: 100293. https://doi.org/10.1016/j.health.2023.100293.

Sammali, Federica, Celine Blank, Tom G. H. Bakkes, Yizhou Huang, Chiara Rabotti, Benedictus C. Schoot, and Massimo Mischi (2021), Multi-Modal Uterine-Activity Measurements for Prediction of Embryo Implantation by Machine Learning. *IEEE Access* 9: 47096–47111. https://doi.org/10.1109/ACCESS.2021.3067716.

Shah, Afshan, Syed Saud Naqvi, Khuram Naveed, Nema Salem, Mohammad A. U. Khan, and Khurram S. Alimgeer (2021), Automated Diagnosis of Leukemia: A Comprehensive Review. *IEEE Access* 9: 132097–132124.

Sinaga, Haripin Togap, Marni Siregar, Berlin Sitanggang, and Mincu Manalu (2023), Modified Length Board is Effectively Detecting Stunted Children at Posyandu: A Precision and Accuracy Test. *World Journal of Advanced Research and Reviews* 20(2): 1147–1156.

Singh, Utkrisht, Mahendra Kumar Gourisaria, and Brojo Kishore Mishra (2022), A Dual Dataset Approach for the Diagnosis of Hepatitis C Virus Using Machine Learning. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 1–6. IEEE. https://doi.org/10.1109/CONECCT55679.2022.9865758.

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).

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.

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).

Wang, Jing, Fengmin Wang, Yiran Zhao, and Houbao Xu (2023), A Temporal Continuity Clustering Algorithm for Spatial-Temporal Data Based on GMM and KNN. *IET Conference Proceedings* (2023): 685–689. https://doi.org/10.1049/icp.2023.1714.