STUDY AND ANALYSIS OF MACHINE LEARNING FOR IDENTIFYING HEMATOLOGICAL CANCER

M.NANTHINI
Ph.D Research Scholar
Department of Computer Science
Bharathiar University
Coimbatore – 641046
nandhudass12@gmail.com

Dr.R. PORKODI
Professor
Department of Computer Science
Bharathiar University
Coimbatore – 641046
porkodi_r76@buc.edu.in

***Abstract*—Mutations in DNA will change the function of blood cells and that will root for creating blood cancer disease in homo sapiens. It is also known as hematological cancer that affect blood and bone marrow. Diagnosing the disease in early stage and provide accurate treatment in timely manner helps to medical experts for decision making. The main objective of this research is to develop the classification model for identifying the three types of blood cancer such as Leukemia, Lymphoma and Myeloma through biological sequences as input using machine and ensemble learning algorithms. From the experiments results, the performance of the classifiers was assessed and report that decision tree and bagging models achieved effective results for hematological cancer prediction than other algorithms.**

Keywords- *Blood cancer, Hematological cancer, Classifier, Machine learning and Ensemble learning.*

I. INTRODUCTION

Blood cancer, also known as hematologic cancer, encompasses a range of malignant conditions that affect the blood, bone marrow, and lymphatic system. This group of cancers disrupts the production and function of blood cells, which play a critical role in the immune system and overall bodily function. Leukemia is Unveiling the Malignancy of Blood-Forming Tissues and it is a type of blood cancer characterized by the uncontrolled proliferation of abnormal white blood cells in the bone marrow, leading to the disruption of normal hematopoiesis. This condition can be classified into various subtypes, including acute and chronic forms, each with unique clinical presentations and prognostic implications. Acute leukemia, such as Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), typically progresses rapidly and requires immediate intervention. In contrast, chronic leukemias, like Chronic Lymphocytic Leukemia (CLL) and Chronic Myeloid Leukemia (CML), progress more slowly and may remain asymptomatic for extended periods. The study of leukemia has significantly advanced our understanding of cancer biology, particularly in areas related to genetic mutations, cellular signaling pathways, and the microenvironment of bone marrow. Research in leukemia has also paved the way for targeted therapies and personalized medicine, offering new avenues for effective treatment and management.

Lymphoma is the Complex Landscape of Lymphatic Malignancies and it is a group of blood cancers that originate in the lymphatic system, a critical component of the immune system responsible for fighting infections and maintaining fluid balance in the body. Lymphomas are broadly categorized into Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL), each with distinct histological features and clinical behaviors. Hodgkin lymphoma is characterized by the presence of Reed-Sternberg cells, whereas non-Hodgkin lymphoma comprises a diverse group of malignancies with varied origins and molecular profiles. The etiology of lymphoma involves complex interactions between genetic predispositions, environmental factors, and infectious agents. Advancements in molecular biology and immunology have led to the identification of specific biomarkers and therapeutic targets, revolutionizing the approach to lymphoma diagnosis and treatment. Current research is focused on understanding the mechanisms of lymphoma pathogenesis, developing novel immunotherapies, and exploring the potential of combination treatments to improve patient survival rates.

Myeloma (Insights into Plasma Cell Disorders) Multiple myeloma is a type of blood cancer that affects plasma cells, a subset of white blood cells responsible for producing antibodies to combat infections. This malignancy leads to the excessive accumulation of abnormal plasma cells in the bone marrow, resulting in bone damage, anemia, kidney dysfunction, and immune suppression. Myeloma is characterized by the presence of monoclonal protein (M protein) in the blood or urine, which serves as a crucial diagnostic marker. The pathophysiology of myeloma involves genetic abnormalities, such as chromosomal translocations and mutations, which contribute to disease progression and resistance to treatment. Recent advancements in proteomics and genomics have facilitated the identification of key molecular pathways involved in myeloma, enabling the development of targeted therapies and novel treatment strategies. Current research efforts are directed towards understanding the clonal evolution of myeloma cells, optimizing therapeutic regimens, and exploring the potential of immunomodulatory drugs and bone-targeted therapies to enhance patient outcomes.

The integration of cutting-edge technologies, such as next-generation sequencing, single-cell analysis, and CRISPR gene editing, has revolutionized blood cancer research, paving the way for personalized medicine approaches and innovative therapeutic interventions. Immunotherapy, including CAR T-cell therapy and monoclonal antibodies, has emerged as a promising treatment modality, providing durable responses in patients with relapsed or refractory blood cancers. Furthermore, the exploration of the tumor microenvironment and its role in disease progression has opened new avenues for targeted therapies aimed at disrupting cancer cell interactions with their surrounding milieu. Future research endeavors are focused on overcoming therapeutic resistance, improving early detection methods, and enhancing our understanding of cancer biology to develop effective strategies for blood cancer prevention, diagnosis, and treatment.

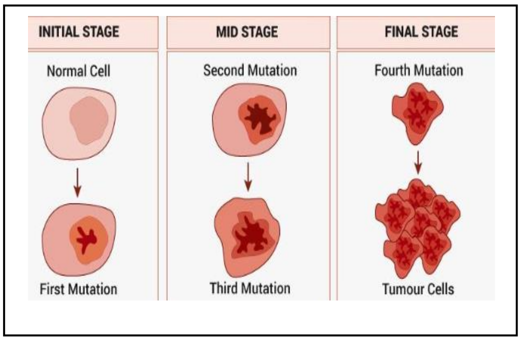


Fig.1 Hematological *Cancer* Life Cycle

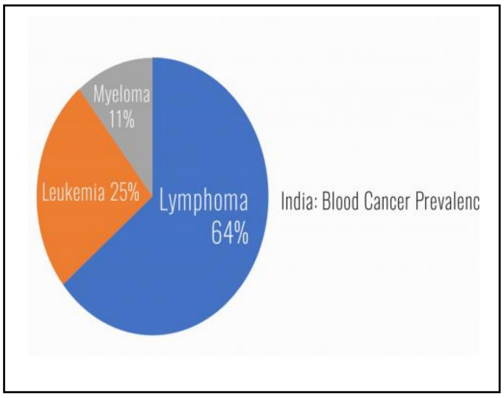


Fig.2 Statistics Of Blood Cancer Types

Information about the life cycle and prevalence of hematological cancer is conveyed by Fig 1 and Fig 2. Blood cancers can manifest in various forms, including leukemia, lymphoma, and myeloma, each with distinct pathophysiological characteristics and clinical manifestations. Understanding these cancers is crucial due to their prevalence and impact on global health. Recent advances in medical research have led to the development of innovative diagnostic and therapeutic approaches, offering hope for improved patient outcomes. This paper aims to provide an overview of blood cancer, focusing on leukemia, lymphoma, and myeloma, and explores their clinical significance, treatment modalities, and the current state of research in these fields.The study of blood cancers, including leukemia, lymphoma, and myeloma offering insights into the molecular mechanisms underlying cancer development and progression.

II. LITERATURE SURVEY

In 2023, Fatma M. Talaaat and Samah A. Gamel developed a classification model that detect leukemia-free and leukemia-affected images. The dataset was collected from Kaggle. The model was developed using CNN with fuzzy optimization and achieved 99.99% accuracy. [1]

In 2021, N.Saranya, N.Kanthimathi, P.Ramya, N.Kowsalya and S.Mohanapriya proposed image processing technique for calculating number of RBCs, WBCs and platelets and also identify the person having leukemia or not. [2]

In 2024, Ebtisam AbdullahAlabdulqader, AishaAhmedAlarfaj, Muhammad Umer, Ala Abdulmajid Eshmawi, ShtwaiAlsubai,Tai-hoon Kim, ImranAshraf proposed improving prediction of blood cancer using leukemia microarray gene data and chi2 featureswith weighted convolutional neural network. They developed a supervised machine learning algorithm CNN for blood cancer identification. A corpus of 22,283 gene leukemia microarray gene data were utilized. A model obtained 99.9% accuracy in this work.[3]

In 2017, Asem H. Shurrab; Ashraf Y. A. Maghari proposed a concept for blood diseases detection using data mining techniques. Decision Tree, Rule Induction, and Naïve Bayes algorithms were used. The blood dataset collected from Europe Gaza Hospital, Gaza Strip. It was observed that Rule Induction, and Naïve Bayes classifiers obtained better results when compared to Decision tree to predict the three types of blood diseases (Hematology Adult, Hematology Children and Tumor).[4]

In 2020, Samir Emad Labib, Christina Albert Rayed proposed a methodology for developing prediction model for childhood leukemia based on data mining classification algorithms. It was observed that the decision tree model is fit for childhood leukemia prediction and obtained the highest accuracy than naïve bayes and random forest. [5]

In 2022, N.D. Jambhekar and P.S. Joshi implemented a model for Detection of White Blood Cell Cancer Diseases through classification techniques. Random forests, support vector machines, closest neighbour algorithms, neural network techniques, and decision tree algorithms were used in this research.[6]

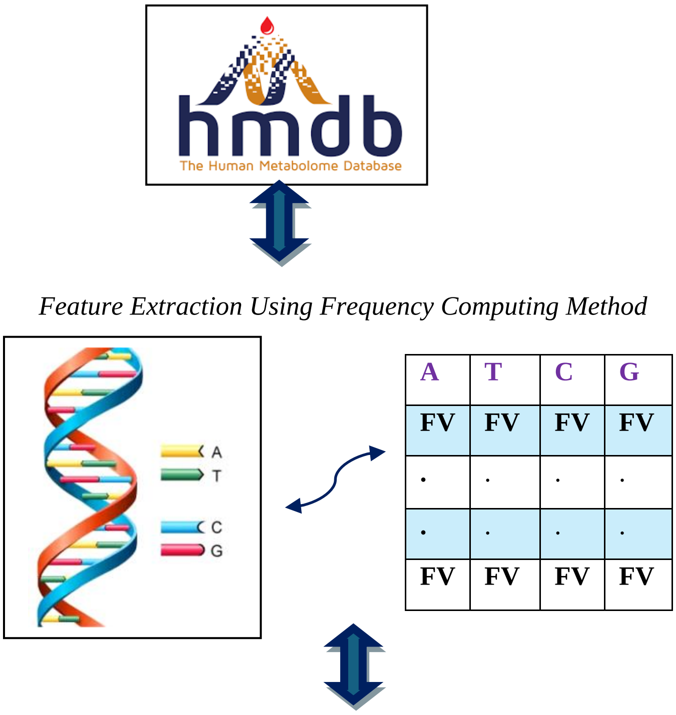
From the analysis of survey, it is observed that blood or hematological cancer detection was performed using machine learning techniques with image or gene data. Hence it is motivated that hematological cancer can also be done through DNA sequences with machine and ensemble learning methods.

III. PROPOSED WORK

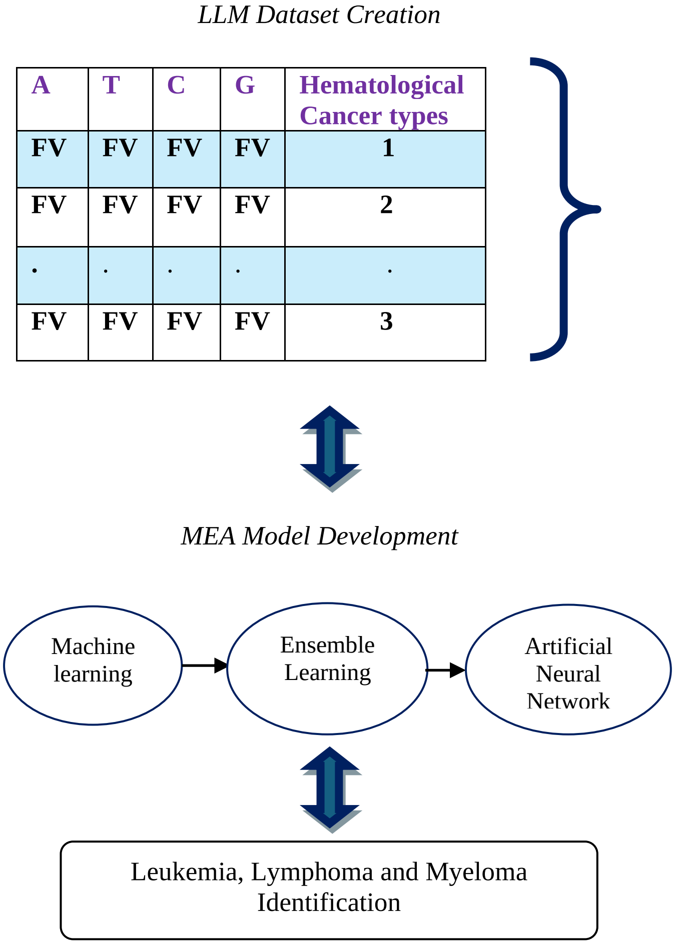
The goal of the proposed work is to develop a classification model for identifying the hematological cancer types using machine learning techniques with contribute features. The first step of this research begins with data collection. Data collection involves gathering information with targeted variables to build a dataset that will be used to train, validate, and test a machine learning model.

The efficiency of the dataset depends on two factors, namely error-free data and its feature values, which are used to train, test, and evaluate a machine learning model. The model development is a crucial phase, and its interpretability can be enhanced through features that are well-engineered, leading to a better understanding of the predictions of machine learning problems.

In this work, the biological sequences of leukemia, lymphoma, and myeloma are collected from the NCBI database. The feature vectors are obtained using the frequency computing method. It calculates the occurrence of each base, namely A, G, C,T and the total length of the sequence. Hematological cancer category detection is implemented through model development using various classification algorithms. This work is carried out in three phases. The first phase of work is to classify the blood cancer types using machine learning classifiers such as Decision Tree, KNN, Naive Bayes, and SVM. The ensemble learning algorithms, namely bagging, gradient boosting, and voting, are used in the second phase. The third phase is demonstrated through an artificial neural network called RBFN (radial basis function networks). The building blocks of the proposed work are organized as follows: In sections 1 and 2, data collection and feature extraction are illustrated. Similarly, data set creation and model implementation are described in Section 3 and Section 4. The entire work flow of this research is shown in Fig. 3.


hmdb
The Human Metabolome Database
Feature Extraction Using Frequency Computing Method
A
T
C
G

A	T	C	G
FV	FV	FV	FV
.	.	.	.
.	.	.	.
FV	FV	FV	FV



A	T	C	G	Hematological Cancer types
FV	FV	FV	FV	1
FV	FV	FV	FV	2
.	.	.	.	.
FV	FV	FV	FV	3

Fig.3 Hematological Cancer Types Identification Model

A. DATA COLLECTION

DNA (deoxyribonucleic acid) is the molecule that composed of four chemical bases that play a crucial role in encoding the information needed for the growth, development, and functioning of an organism. There are four types of DNA nucleotides namely adenine, thymine, guanine and cytosine and represented by A,T,G,C. The DNA sequence of Hematological cancer types such as leukemia, lymphoma and myeloma are taken from NCBI database. The sample sequence is shown below.

i. Sample sequence for leukemia

GGTAGAGACAGGGTTTCGCCATGTTGGCCACAC
TGCTCTCAAACTCCTGACCTCAAGTGATCCGCC
CGCCATGGTCTCCCAAAGCGCTGCGAGTACGGC
AGGAGCCACCGTGTCCAGCCCAAAATGATTAA
ATGTTACCCAATAGGTCTGCCAACAATCGAAAG
AGAAGCAATCCCATCTCTTTTTCACTGTCTATTT
TTAGGGCCAAAATATATGTGTTGGCAAAAATGC
TGTGGCATATACAAATGGCACAATTCCAAAATC
AACTACTCAACTCTGCTCAAAAGATGTAATTCC
TGATGAAAAATGCAAGCATGACTTCTACCTGAG
CCCAAGGTAATTTTCTTAC

ii. Sample sequence for lymphoma

```
TCAGAATAAACAGACAACCCACAGAA
TGTGAGAAAATATTTGCAAATTATGCA
TCTGACAAAGGTCTAATACCCAGCAAT
CTATAAGGAACTCAAACAAATTAGCA
AGAAAAAAAATCCCATGAAAAGGTAG
ACAAATGACATGAATAGACACTTCTCA
AAATAAGATATATAAATAGCCACAAA
CATATGAAAAAATAATCAACATCACTA
ATCATCAGGTAAATGCAAATTAAAACC
ATAATGAGATACCACCTTATCCCAGCC
AGAATGGCCATTATTAGAAAGTCCAAA
AACAATAGATGTTGGCATGGATGTGGT
GAAAAGGGAAGAGTTTACACTGCGGG
CAGGAATGTAAATTAGTCCAACCTCTA
TGGAGAGCAGTATGCAGATTTCTTTTT
TCTTTTTTTTTTTATTTCGA
```

iii. Sample sequence for Myeloma

```
TAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAA
CCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCT
AACC
```

Table.1 Counts of DNA sequences

Sequence Name	No of sequences
Leukemia	38
Lymphoma	23
Myeloma	39

The table 1 describe the total count sequences of each blood cancer type used in this research. For Leukemia, 38 sequences are collected. For other two types namely Lymphoma and Myeloma, 23 and 39 sequences are taken in identifying the hematological cancer types.

B. FEATURE EXTRACTION

Feature extraction plays important role in improving the performance of models. In this research, DNA sequences are converted into feature vectors through frequency computing method which is explained below.

Step1: Enter the DNA Sequence
Srep2: Calculate the length of sequence
Step3: Count the adenine, thymine, guanine and cytosine Using DNA.count() function
Step4: Compute the frequency for four types of DNA Nucleotides using below formula.

$$\textbf{(i) Frequency of adenine} = \frac{\textbf{Total Count of adenine}}{\textbf{Total length of ACGT}}$$

$$\textbf{(ii) Frequency of thymine} = \frac{\textbf{Total Count of thymine}}{\textbf{Total length of ACGT}}$$

$$\textbf{(iii) Frequency of guanine} = \frac{\textbf{Total Count of guanine}}{\textbf{Total length of ACGT}}$$

$$\textbf{(iv) Frequency of cytosine} = \frac{\textbf{Total Count of cytosine}}{\textbf{Total length of ACGT}}$$

(i) Frequency computing results of leukemia

```
Enter DNA sequence
GGTAGAGACAGGGTTTCGCCATGTTGGCCACACT
GCTCTCAAACTCCTGACCTCAAGTGATCCGCCCGC
CATGGTCTCCCAAAGCGCTGCGAGTACAGGCAGG
AGCCACCGTGTCCAGCCCAAAATGATTAAATGTT
ACCCAATAGGTCTGCCAACAATCGAAAGAGAAGC
AATCCCATCTCTTTTTCACTGTCTATTTTTAGGGCC
AAAATATATGTGTTGGCAAAAATGCTGTGGCATA
TACAAATGGCACAATTCCAAAATCAACTACTCAA
CTCTGCTCAAAAGATGTAATTCCTGATGAAAAAT
GCAAGCATGACTTCTACCTGAGCCCAAGGTAATT
TTCTTAC

29.859
24.507
25.352
18.873
```

(ii) Frequency computing results of lymphoma

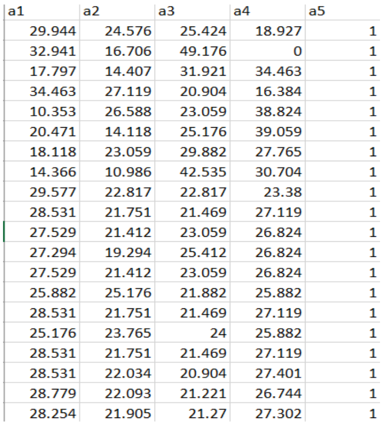
```
Enter DNA sequence
TCAGAATAAACAGACAACCCACAGAATGTGAG
AAAATATTTGCAAATTATGCATCTGACAAAGGT
CTAATACCCAGCAATCTATAAGGAACTCAAAC
AAATTAGCAAGAAAAAAAATCCCATGAAAAGG
TAGACAAATGACATGAATAGACACTTCTCAAA
ATAAGATATATAAATAGCCACAAACATATGAA
AAAATAATCAACATCACTAATCATCAGGTAAAT
GCAAATTAAAACCATAATGAGATACCACCTTAT
CCCAGCCAGAATGGCCATTATTAGAAAGTCCA
AAAACAATAGATGTTGGCATGGATGTGGTGAA
AAGGGAAGAGTTTACACTGCGGGCAGGAATGT
AAATTAGTCCAACCTCTATGGAGAGCAGTATGC
AGATTTCTTTTTTCTTTTTTTTTTTATTTCGA

41.568
24.941
17.102
16.152
```

(iii) Frequency computing results of myeloma

```
Enter DNA sequence
TAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACC
32.998
16.499
49.095
0.0
```

C. DATASET

The LLM dataset contains 102 instances and 4 features. In this work, multilabel classification is used for identifying the three types of blood cancer. The class labels are leukemia, lymphoma and myeloma which is represented as 1,2,3 in the dataset. The dataset is divided into two parts for training and testing the model. Here 80% of data is used for training and 20% of data is used for testing stage. The frequency adenine, thymine, guanine and cytosine is represented by a1,a2,a3 and a4 and The class label is a5.



a1	a2	a3	a4	a5
29.944	24.576	25.424	18.927	1
32.941	16.706	49.176	0	1
17.797	14.407	31.921	34.463	1
34.463	27.119	20.904	16.384	1
10.353	26.588	23.059	38.824	1
20.471	14.118	25.176	39.059	1
18.118	23.059	29.882	27.765	1
14.366	10.986	42.535	30.704	1
29.577	22.817	22.817	23.38	1
28.531	21.751	21.469	27.119	1
27.529	21.412	23.059	26.824	1
27.294	19.294	25.412	26.824	1
27.529	21.412	23.059	26.824	1
25.882	25.176	21.882	25.882	1
28.531	21.751	21.469	27.119	1
25.176	23.765	24	25.882	1
28.531	21.751	21.469	27.119	1
28.531	22.034	20.904	27.401	1
28.779	22.093	21.221	26.744	1
28.254	21.905	21.27	27.302	1

D. MODEL DEVELOPMENT

Automating tasks, handling large datasets, and making data-driven decisions can be supported by machine learning. Ensemble learning enhances the above benefits by improving accuracy, robustness, and performance through the combination of multiple models. Hematological cancer identification model is developed using machine learning algorithms such as Decision tree, KNN, Naïve bayes and SVM and ensemble learning algorithms namely Bagging, Gradient Boosting, Voting and Radial basis function networks. The models are implemented using spyder platform and python language. The MEA model working process is shown below in Fig.4.

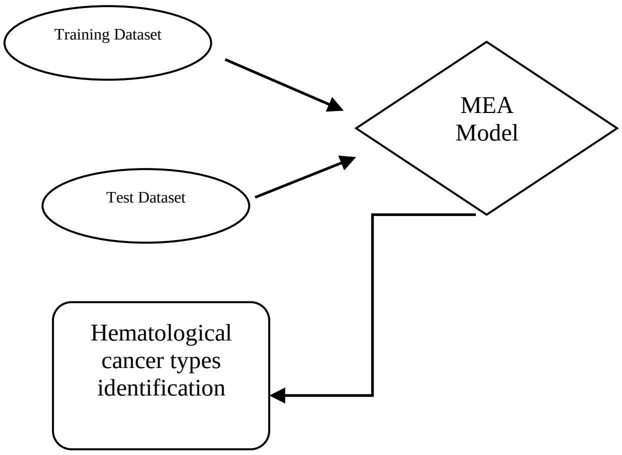


Fig.4. Working flow of MEA model

IV. EXPERIMENTS AND RESULTS

The machine and ensemble classifiers are evaluated for their efficiency using test dataset. The metrics such as precision, recall, F1-score and accuracy are used to assess classifier's output quality and shown in below table II.

Precision is the percentage of correct positive predictions out of all positive predictions made.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{False Positives (FP)+True Positives (TP)}}$$

Recall is the percentage of actual positives that were correctly identified by the model.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{False Negatives (FN)+True Positives (TP)}}$$

The F1 Score balances Precision and Recall.

$$\text{F1 Score} = \frac{2 \times \text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}}$$

Table.2 Performance Measures of Machine Learning Classifier

S. NO	ALGORITHMS	PRECISION	RE-CALL	F1-SCORE	ACCURACY
1	DECISION TREE	1.0	1.0	1.0	100%
2	KNN	0.88	0.78	0.80	80%
3	NAÏVE BAYES	0.47	0.48	0.47	52%
4	SVM	0.96	0.95	0.95	95%
5	BAGGING	1.0	1.0	1.0	100%
6	GRADIENT BOOSTING	0.96	0.95	0.95	95%
7	VOTING	0.93	0.81	0.90	90%
8	RADIAL BASIS FUNCTION NETWORK	0.95	0.95	0.94	0.95

A. COMPARATIVE ANALYSIS

The performances of all the classification models are compared against their evaluation metrics. From comparative analysis of machine learning, it is observed that decision tree classifier achieves effective results than other algorithms for recognizing hematological cancers. By analyzing the results of ensemble learning classifiers, it is perceived that the bagging model shows high accuracy than other classifiers. The comparative analysis of MEA models is shown below in Fig.5.

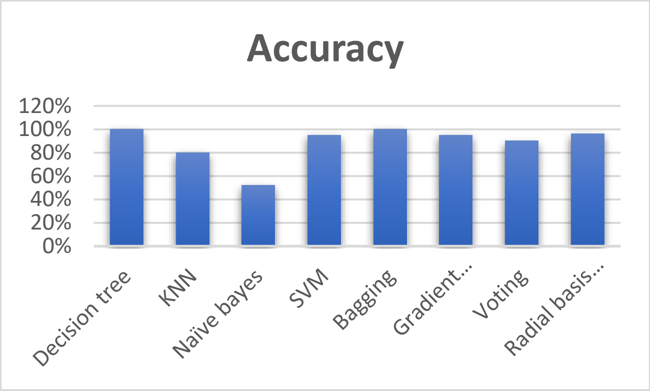


Fig.5.Accuracy analysis

V. CONCLUSION

The discovery of hidden patterns or insights from medical data and the identification and development of vaccines for diseases is an intriguing research problem. This paper demonstrates the implementation of Hematological cancer prediction using machine learning and ensemble learning algorithms. The DNA sequences were collected from NCBI database. The features are transformed into frequency vectors, and the LLM dataset is created which given as an input to the various classifiers. The results are analyzed and report that decision tree and bagging classifiers is suitable for hematological cancer identification. The future of this research work can be enhanced by adding more instances and features to improve the model performance and use advanced learning techniques.

References

[1] Fatma M. Talaaat and Samah A. Gamel, "Machine learning in detection and classification of leukemia using C-NMC_Leukemia", 2023.

[2] N.Saranya, N.Kanthimathi, P.Ramya, N.Kowsalya and S.Mohanapriya,,"Blood cancer detection using machine learning", 2021.

[3] Ebtisam AbdullahAlabdulqader, AishaAhmedAlarfaj, Muhammad Umer, Ala Abdulmajid Eshmawi, ShtwaiAlsubai,Tai hoon Kim, ImranAshraf , "Improving prediction of blood cancer using leukemia microarray gene data and chi2 features with weighted convolutional neural network",2024.

[4] Asem H. Shurrab; Ashraf Y. A. Maghari, "Blood diseases detection using data mining techniques",2017

[5] Samir Emad Labib, Christina Albert Rayed," Prediction Model for Risk Factors of Childhood leukemia Based on Data Mining Classification algorithms.",2020.

[6] N.D. Jambhekar and P.S. Joshi," Detection Of White Blood Cell Cancer Diseases Through Classification Techniques",2022.

[7] S.Neelamegam, Dr.E.Ramaraj, "Classification algorithm in Data mining: An Overview", International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 Issue 8- Sep 2013, ISSN: 2249-2615.

[8] Han, J. and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Publishers, 2006.

[9] T. Swarna Latha, "Recognition of Blood Cancer Using Different Classification Techniques",2022.

[10] Baker Khalid Baker, Rakan Mohammed Rashid, Nashat Salih Abdulkarim Alsandi, Omar Farook Mohammad," Classification of Image Blood Cancer by Using Multi-Training RNN",2021.

[11] Asad Ullah , Muhammad Shoaib, " Normal Versus Malignant Cell Classification in B-allwhite Blood Cancer Microscopic Images Using Deep Learning", 2022.

[12] Mukesh Madanan, Anita Venugopal, and Nitha C.Velayudhan," Designing an Artificial Intelligence Model using Machine Learning Algorithms and Applying it to Hematology for the Detection and Classification of Various Stages of Blood Cancer", 2020.

[13] Pradeep Kumar Das, Sukadev Meher, " An Efficient Deep Convolutional Neural Network Based Detection and Classification of Acute Lymphoblastic Leukemia, 2021.

[14] D. Umamaheswari, S.Geetha, "A framework for efficient recognition and classification of acute lymphoblastic leukemia with a novel customized-KNN classifier", 2018.

[15] Sumit Kumar Das, Kazi Soumik Islam, Tanzila Ashsan Neha, Mohammad Monirujjaman, Sami Bourouis, Towards the Segmentation and Classification of White Blood Cell Cancer Using Hybrid Mask-Recurrent Neural Network and Transfer Learning", 2021.