e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711056

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Data-Driven Disease Prediction and Classification in Biomedical Informatics

Mrs Deepika K S, Ajay Kumar D, Tejas A

Assistant Professor, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

ABSTRACT: When a lot of immature lymphocyte blood cells multiply and prevent other blood cells from growing, leukemia, a type of cancer, develops. The Complete Blood Count (CBC) test is commonly used to identify leukemia. In the complete blood count (CBC), leukemia is indicated by either an excess or a deficit of a particular type of blood cell. Despite affecting children 80% of the time, leukemia can be fatal in adults. Therefore, early leukemia detection is crucial for life preservation. Rapid leukemia diagnosis is possible with the aid of computer technologies. Then, utilizing the features, many machine learning classification methods were trained and evaluated, including SVM and Random Forest. By evaluating its performance in relation to multiple performance metrics, including accuracy and precision, the best classifier can be identified.

KEYWORDS: Confusion matrix, classification, algorithms, data mining, accuracy, and evaluation.

I. INTRODUCTION

The body as a whole receives vital nutrients from the blood.
Leukocytes, thrombocytes, white blood cells, and platelets are the three primary blood cell types produced by the human body. Erythrocytes are the red blood cells. [1] Every cell in the body receives oxygen from red blood cells (RBCs). In the event of an injury, platelets help the blood clot. White blood cells, or WBCs, provide protection to the human body against infections and illness. White blood cells only make about 1% of blood, yet being vital to the human immune system. Even minor adjustments have a significant impact. White blood cells (WBCs) in blood plasma can increase or decrease to signal an illness. Monocytes, neutrophils, eosinophils, basophils, and lymphocytes are the five different types of white blood cells. If the white blood cell count changes, there should be concern expressed. Most often, leukemia is associated with a low White WBC (white blood cell) count. [2] When WBC counts are excessively high, human beings could become disoriented and contribute to illness. Leukemia is a potentially lethal disease and a type of cancer. It harms the bone marrow and blood because to the aberrant white blood cells' rapid multiplication. The production of platelets and red blood cells by bone marrow is hampered by these abnormal WBCs, which also cannot stop illness.

Leukemia can be classified as either acute or chronic.
The symptoms of acute leukemia are more severe and the disease advances more quickly than chronic leukemia. Myelogenous and lymphocytic leukemia are the two types.

The immune system is supported by white blood cells known as lymphocytes. [3] The excessive proliferation of malignancies in the bone marrow cells that produce those lymphocytes is known as lymphhocytic leukemia. The hematopoietic cells that produce RBCs, WBCs, and platelets multiply aberrantly in myelogenous leukemia. The four disease types that make up leukemia are ALL, AML, CLL, and CML (chronic lymphocytic leukemia).

Children are most commonly affected by acute lymphoblastic leukemia (ALL), a kind of cancer. The majority of ALL cases involve healthy people, although a very small percentage of patients have inherited traits such environmental factors or familial risk. [7] It is distinct due to its distinctive chromosome abnormalities and genetic lymphoid changes. Although ALL is responsible for 80% of juvenile leukemia, it only accounts for 20% of adult leukemia.

© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711056

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

ALL is diagnosed using complex and time-consuming procedures. Since the advent of risk-adapted medications and supportive care, the survival rate in developed countries has risen to 90%.Data classification is a frequent problem in data mining. The goal of the computer science field of data learning is to create methods that enable computers to learn. It serves as a guide for the kind of patterns that must be discovered throughout the data mining procedure [4]. Data mining models can be either descriptive or predictive, and their goal is to identify the model that best fits the data under study. Data mining is the procedure used to extract information from a sizable collection of data. Finding leukemia early is essential to saving lives.

Rapid leukemia diagnosis can be aided by computer technology. [10] Then, utilizing the parameters, several machine learning classification methods, including SVM and Random Forest, were trained and evaluated. It is possible to identify the most efficient classifier by assessing its performance in relation to multiple performance criteria, such as accuracy and precision. methodologies are available. under the present circumstance. We've discovered that it employs many methods.Every technique offers intriguing insights on a cyberattack [6]. According to several research, they make the cyberattack model far more effective. Nevertheless, different models cannot compromise the most recent security system. We have put forth a plan for creating cyberattacks in this work. We have developed a real-time assault model design environment.

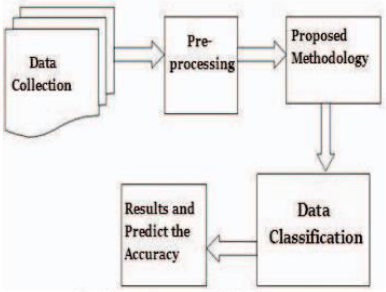

Fig. 1. :Proposed Workflow

II. MATERIALS AND METHODS

A. Leukemia

A. Leukemia

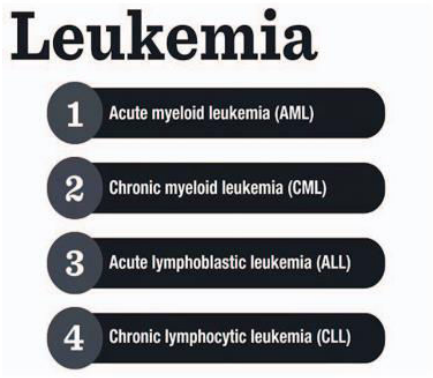

Fig. 2. Types of Leukemia

The bone marrow is where leukemia, also known as blood cancer, starts and develops a significant number of aberrant cells. ALL, AML, CLL, and CML (chronic lymphocytic leukemia) are the four disease forms that comprise leukemia. [11] Another name for acute lymphocytic leukemia is acute myelogenous leukemia. Figure 2 lists the many types of leukemia.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Computational Techniques

B. Computational Techniques

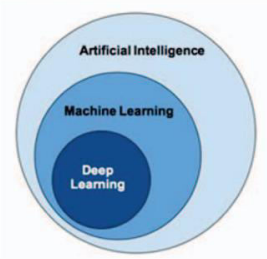

Fig. 3. Computational Techniques

Cancer can be predicted using a variety of computer techniques. Figure 3 [5] presents a set of computationally efficient methods. In artificial intelligence, deep learning is a subset of machine learning and a superset of it. [12].
Artificial Intelligence is an area in computer science.A computer with AI capabilities can complete a task more intelligently. The machine can become smarter with AI. AI machine input necessitates a process of reasoning with its actions. All things that humans can complete can be completed by AI.

There is a chance that AI will boost human productivity. AI has a number of drawbacks as well. The two types of AI that are commonly found are weak and strong AI.

Weak AI: Occasionally It is known as "Narrow A." Weak AI only works in very restricted situations and is very good at just one task at a time. A human has preprogrammed every action. Google search, image recognition, Siri, Alexa, and so forth are a few examples.

Strong AI: This kind of AI is unquestionably a more intelligent machine. It has intelligence and operates in multiple dimensions. It is capable of processing the sentimental statement and using the input to produce a strong choice. Examples include chatbots, social media surveillance tools, chessboard games, drone robots, and personalized healthcare treatment recommendations.

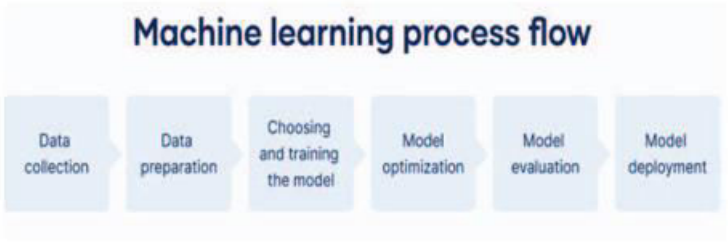

Fig. 4. ML Process

Without being specifically programmed, it gives the system the capacity to learn on its own and get better with time. As illustrated in figure 4, machine learning methodologies begin with preprocessing and feature engineering, then go on to model development and classification. How successful these strategies are is greatly influenced by the attributes that are selected, which may or may not be the optimal traits for class discrimination.

The initial step in the machine learning process is gathering the dataset. Use machine learning to create a model for the system after using feature learning to identify key traits. Ultimately, the model classifies the class label. Two forms of machine learning are supervised and unsupervised learning [6].

One kind of categorization method that involves gathering a sample of labeled data is called supervised learning. When training the model, this supervised machine learning finds patterns in the data so that the machine may learn from a

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

sample of data. When an unknown sample of data is later entered into the model, it detects the test sample and, using a probabilistic value, places it into one of several categories. [9] The model receives labeled data, is trained using machine learning techniques, and, once test data samples are supplied, predictions are generated.

Lastly, the sample of provided data can be classified using the model.
Unsupervised machine learning is another type of categorization technique that does not label input samples.
It signifies that they are not categorized and that there are no data samples that match any specific class of the categories. After being given the training sample data set, the machine learning model uses methods like k-Means clustering and association rules to uncover the hidden pattern in the input data. An unsupervised machine learning system can classify the data according to the similarities among its attributes. [15] In the second stage, the model is fed the unknown data sample as test data sets in order to find the characteristics using the current model and subsequently categorize the data.

ML algorithms evaluate unlabeled input data, and during processing, an unsupervised machine learning logarithm can categorize the data sample according to the.

C. Data Processing for Leukemia in Healthcare associated illnesses
As digital technology advances, it has led to the increase in data in the last few years. The process of digitizing patient data, Simple embedded systems in healthcare systems access to fast mobile networks, as well as evolving
All of these technologies are accelerating the growth of data in applications in healthcare. reports on clinical and diagnostic procedures, prescription drugs, health records, and computerized health insurance coverage, wellness reports, and Information about health in medical journals helps to Big Data in healthcare [13]. manually examining and analyzing vast amounts of complex data, referred as asFor academics, the Big Data problem becomes a challenging task.

Big data management is essential to the healthcare sector.Big data analytics greatly aids patients and doctors in providing effective care in the healthcare industry. The four obstacles of processing big data are diversity, velocity, volume, and veracity. Data collection yields enormous amounts of information, especially for leukemia patients. Making inferences from the enormous amount of available healthcare data requires methods and procedures that can manage such enormous volumes of data in a secure and efficient manner.

D.Utilizing machine learning to process massive volumes of data.
The machine learning methods replace the automatic learning from a data collection. It starts with gathering data and using it to make predictions. Depending on the level of expertise, machine learning models can be classified as supervised, semi-supervised, or unsupervised. Compared to a doctor, a machine learning program might go at patient records for a lot more information. With additional features, the model's ability to recognize patients with health issues was much enhanced.

E. Data Mining's Significance in Health Informatics
Numerous disciplines, including machine learning, neural networks, statistics, and pattern recognition, are combined in data mining. Its primary focus is on the process of computationally extracting latent knowledge structures from large data repositories represented by models and patterns. The practice of medicine depends on data. Every second, a significant number of running processes produce new data. The advancement of computers and new algorithms has led to a boom in computer tools for the healthcare industry, which can no longer ignore these emerging technology. Consequently, the integration of computing and healthcare led to the creation of bioinformatics.

It is one of the areas of the health sector with the fastest growth rates and encompasses a wide spectrum of research and applications. It works with biomedical data, information, and knowledge.

Intelligent algorithms and machine learning can be used in decision-making and problem-solving systems to provide high-quality healthcare [13].

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

F. Data mining applications in medical applications for healthcare

One of the most significant sectors of industrial civilizations is health and medicine. The data mining technique can help find the rules controlling the onset and spread of epidemic diseases by learning from vast amounts of disease and medical record data. [8] The popular medical programs include,

Health care service cost prediction.

Determination of disease treatment

Prognosis and diagnosis of diseases

The scientific topic of health informatics is one that is expanding quickly. It deals with collecting, storing, retrieving, transmitting, and optimizing health-related data, information, and knowledge. This area of study aims to improve patient care and community health by applying to clinical care, nursing, public health, and biological research.

G. Systems for Supporting Decisions

Knowledge-based systems that support information and facilitate decision-making are referred to as decision supportsystems. Scientific fields. To ascertain the patient's condition and the kind of care required, physicians can input the patient's information into Utilize a decision support system and electronic health forms to Examine the information. Clinical data and language standardization interaction between businesses are two ways to optimize the advantages of data mining technologies for healthcare.[14] Additionally, it derives consistent patterns from biological and databases for healthcare, including connections between diseases and health issues, connections between illnesses, and connections between drugs, etc.

S. No	Statistical Approaches
1	Statistical Analysis: Basic statistical techniques are used for tasks like data summarization, hypothesis testing, and determining the significance of observed patterns in biological data.
2	Regression Analysis: Linear and nonlinear regression models are applied to understand relationships between variables and predict outcomes.
3	Clustering: Methods like k-means clustering and hierarchical clustering are employed for grouping similar biological entities, such as genes or proteins, based on their expression patterns.
4	Classification: Algorithms like logistic regression and support vector machines are used for tasks like gene function prediction and disease classification based on gene expression profiles.
5	Enrichment Analysis: Statistical methods like Gene Set Enrichment Analysis (GSEA) are used to identify overrepresented biological terms in a set of genes or proteins.
6	Hidden Markov Models (HMMs): Applications include sequence alignment, protein family modeling, and gene prediction. HMMs are used to simulate biological sequences and probabilistic changes between hidden states.
7	Ensemble methods, such as gradient boosting and random forests: Application: feature selection and predictive modeling. Ensemble approaches integrate several base models to enhance overall performance and provide good generalization across a range of datasets.

S. No	Deep Learning Approaches
1	Neural Networks: Deep neural networks, including CNNs and RNNs, are applied for tasks like image analysis (e.g., microscopy images of cells), sequence analysis (e.g., DNA, RNA, protein sequences), and predicting biological properties.
2	Deep Generative Models: VAEs and GANs are used for generating synthetic biological data, which can be valuable for training other machine learning models.
3	Deep Transfer Learning: Pretrained DL models (e.g., from NLPs tasks) are fine-tuned for specific bioinformatics tasks when the available biological data is limited.
4	Graph Neural Networks (GNNs): GNNs are used for tasks involving biological networks, such as protein- protein interaction networks and drug-target interaction networks. GNNs can capture complex relationships in graph-structured data.
5	Reinforcement Learning: This techniques are applied in bioinformatics for tasks such as drug discovery, where algorithms learn to make a sequence of decisions to optimize a reward function.
6	CNNs (convolutional neural networks): Application Cell categorization using images and DNA sequence analysis. Due to its design for handling grid-like data, CNNs are well suited for applications involving grid-like structures, such as those involving images and sequences.
7	GANs, or generative adversarial networks: Application: Generating new molecular structures and drug discovery. Two neural networks, a discriminator and a generator, are combined to train a GAN in a competitive environment. GANs can produce artificial data that is comparable to a given dataset.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. OUTCOMES AND TALK

A well-known leukemia dataset is the "ALL-AML dataset," also known as the "Golub Leukemia Dataset." Since Golub et al. first introduced this dataset in 1999, it has been extensively used in several studies. It contains the gene expression profiles of leukemia patients, each of which displays the activity levels of hundreds of patient-specific genes.

In the field of machine learning, the Random Forest approach is widely accepted to be a very effective and popular technique. The aforementioned algorithm is made up of many decision trees, which together use the output of each class to ascertain the output's mode. solitary tree. Several trees are produced once decision trees are constructed using the sample sets, eventually resulting in the creation of a forest. Decision trees are a very useful tool when it comes to classification. problems, including this one, and for other assignments, including regression. The reason for this is because decision trees produce a wide variety of trees throughout the training phase and then generate the classes or average forecasts for each solitary tree. To reduce variance, many deep decision trees are trained on different subsets of the same dataset and then averaged. RF classifies data by combining many DTS. By averaging individual tree predictions to avoid overfitting and capture intricate feature interactions, it enhances generalization. Since they model the relationship between tumor traits and diagnosis in illness categorization, it is a reliable and adaptable approach.

Due to its parameter sensitivity, the technique's effectiveness can be affected by the number of attributes utilized to split a node, tree depth, and ensemble size. It employs fine-tuning since employing techniques like cross-validation to modify parameters is necessary to obtain the best results. Heterogeneous feature types, non-linear correlations, and noise resistance are all handled by RF.

A well-liked supervised machine learning method for pattern identification and classification problems is the Support Vector Machine (SVM), particularly when the dataset contains exactly two classes.

To determine the optimal hyperplane for effectively splitting the different classes, Support Vector Machines (SVMs) are employed. The classifier uses a feature vector, which is an input pattern, to decide its categorization.

The approach may have trouble classifying data if the feature vectors are not linearly separable, but it can effectively categorize data that is linearly separated. According to the literature, the kernel technique has been used to tackle this issue. In order to transform input data into higher dimensional space, Support Vector Machines (SVM), which offer a rapid training process, employ kernel techniques. This method can be applied to regression analysis and pattern classification. For a support vector machine (SVM) classifier to be effective, the kernel function must be chosen carefully.

We employ different kernel functions for different categorization tasks. The Support Vector Machine (SVM) course, which is available via the sci-kit-learn package, was used to finish this project's SVM application. The Support Vector Machines (SVM) may present challenges.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In terms of memory use and could necessitate complex tuning and interpretation.

True Positive (TP):	This represents the number of leukemia cases that are correctly identified as positive by the classifier. In other words, it is the number of leukemia patients correctly classified as having leukemia.
True Negative (TN):	This represents the number of non-leukemia cases that are correctly identified as negative by the classifier. It is the number of healthy individuals correctly classified as not having leukemia.
False Positive (FP):	This represents the number of non-leukemia cases that are incorrectly identified as positive by the classifier. It is the number of healthy individuals falsely classified as having leukemia.
False Negative (FN):	This represents the number of leukemia cases that are incorrectly identified as negative by the classifier. It is the number of leukemia patients falsely classified as not having leukemia.

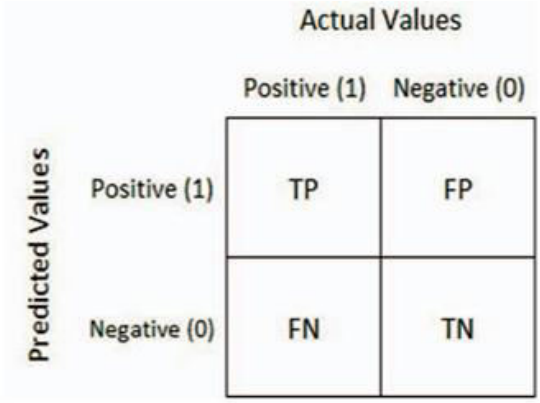

Fig. 5. Confusion Matrix

The matrix construction is shown in Fig. 5.
The hybrid method is a recommended strategy for this study.
This approach integrated RF and SVM algorithms.
The results are displayed graphically in Figures 6, 7, and 8.

© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711056

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

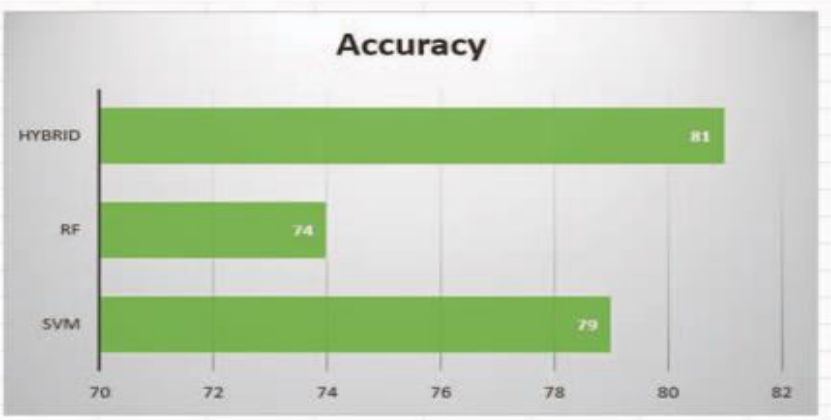

Fig. 6. Accuracy Comparison

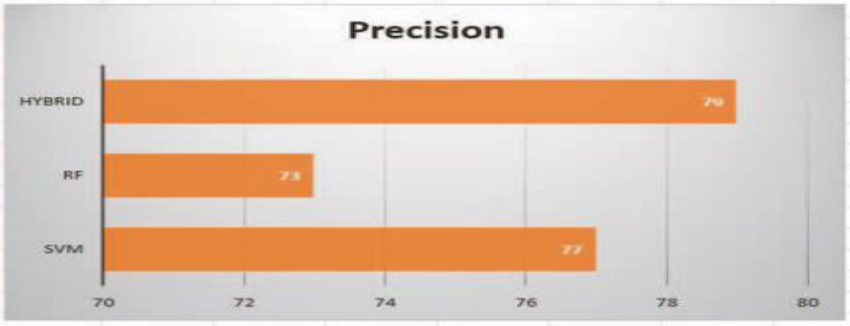

Fig. 7. Precision Comparison

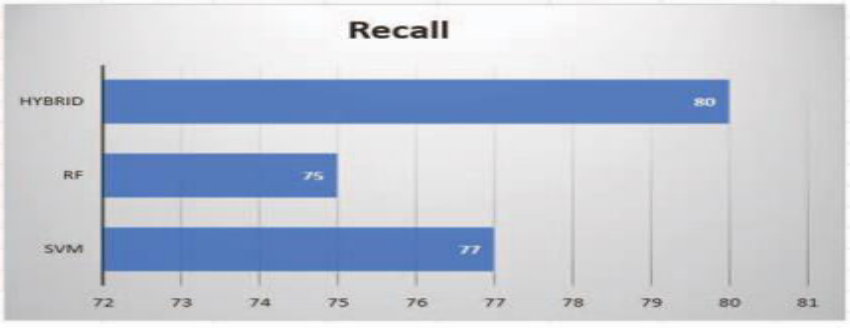

Fig. 8. Recall Comparison

IV. CONCLUSION

The suggested hybrid algorithm improves predicted accuracy over individual models by utilizing the advantages of several machine learning approaches. Group techniques take advantage of the variety of base models, reducing shortcomings and improving performance as a whole. CrossValidation guarantees generalization and resilience, which lowers the danger of overfitting. By combining forecasts from several models, making the hybrid algorithm less vulnerable to biases present in all models. Furthermore, by combining a number of base models, the algorithm's adaptability enables it to be tailored to particular datasets and requirements. This makes it suitable for a variety of data types and problem domains. However, to develop an effective ensemble technique and determine the optimal weights for merging predictions, a great deal of testing and validation are required. When deploying in real healthcare settings,

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

there are additional challenges with data security, regulatory compliance, and interpretability. AI algorithms are powerful diagnostic and analytical tools in the healthcare industry that can solve problems like conflicting expert opinions and laborious manual diagnostics. By enhancing classification models, this work aimed to improve leukemia detection and classification accuracy.

Future studies could examine parallel processing methods to execute many jobs concurrently, saving time, as the existing system only manages one operation at a time.

REFERENCES

[1] Sunita Chand, and V.P. Vishwakarma, “Applications of deep learning in acute leukemia detection-a review,” AIP Conf. Proc. 2782, 020079 (2023).
[2] The article "Development and Evaluation of a Leukemia Diagnosis System Using Deep Learning in Real Clinical Scenarios" was published in Front Pediatrics in 2021.
[3] The 7th International Conference on I-SMAC, 2023, featured an analysis of automated leukemia cancer detection using feature selection and classification techniques by B.Divyapreethi and A. Mohanarathinam.
[4] Bilal Wajid, Sajida Zahid, and at el, “Survival Rate Prediction of Blood Cancer (Leukemia) Patients Using Machine Learning Algorithms,” International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering, 2022.
[5] Machine Learning-based Algorithmic Approach for Leukemia Detection and Classification, Second International Conference on Smart Technologies and Systems for Next Generation Computing, 2023, S. Karthikeyan, E. Thirisha, et al.
[6] International Conference on Innovative Data Communication Technologies and Application, 2023; Amogh Ramagiri, Sathwik Gottipati, et al., "Image Classification for Optimized Prediction of Leukemia Cancer Cells using Machine Learning and Deep Learning Techniques."
[7] Anika Tasnim and Bayezid Islam, “Analysis of Different Feature Selection Techniques on Different Machine Learning Approaches to Classify Leukemia Subclasses,” IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, 2022.
[8] "A Brief Overview of Machine Learning Algorithms," Susmita Ray, International Conference on Machine Learning, Big Data, Cloud, and Parallel Computing, 2019.
[9] Pardeep Kumar, Shuchita Upadhyaya and Kirti Gupta, “Predictive Analysis in Healthcare: A Survey,” Seventh International Conference on Parallel, Distributed and Grid Computing, 2022.
[10] Shruti and Naresh Kumar Trivedi, “Predictive Analytics in Healthcare using Machine Learning,” 14th International Conference on Computing Communication and Networking Technologies, 2023.
[11] In their 2020 paper, "Comparison of Data Mining Techniques in Healthcare Data," Alaa Hussein Al-Hamami, Mustafa Tareq Abd, and Zeki Saeed Tawfik.
[12] Rasha Anwar Mohammed and Khattab M Ali Alheeti, “Intelligent Identification Apparoch for Healthcare Systems,” Fifth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU), 2022.
[13] "Identifying and Predicting Chronic Diseases Using Machine Learning Approach," by Rayan Alanazi, J Healthc Eng. 2022, 2826127.
[14] "A Framework for Early Detection of Acute Lymphoblastic Leukemia and Its Subtypes From Peripheral Blood Smear Images Using Deep Ensemble Learning Technique," by Abdullah Alourani, Muhammad Shahbaz, and others, IEEE Access, Year: 2024 | Volume: 12.
[15] "Hybrid Ant Lion Mutated Ant Colony Optimizer Technique With Particle Swarm Optimization for Leukemia Prediction Using Microarray Gene Data," by A. Balajee, H.S. Shreenidhi, and colleagues, Jonnakuti Rajkumar Annand, IEEE Access, 2024 | Volume: 12.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com