Asian Journal of
Research in
Computer Science

Asian Journal of Research in Computer Science

9(3): 52-65, 2021; Article no.AJRCOS.69216
ISSN: 2581-8260

Leukemia Diagnosis using Machine Learning Classifiers Based on Correlation Attribute Eval Feature Selection

Revella E. A. Armya[1*], Adnan Mohsin Abdulazeez[2], Amira Bibo Sallow[3] and Diyar Qader Zeebaree[4]

[1]Akre Technical College of Informatics, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.
[2]Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.
[3]Nawroz University, Duhok, Kurdistan Region, Iraq.
[4]Research Center of Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

Authors' contributions

This work was carried out in collaboration among all authors. Author REA managed the literature searches related to leukemia diagnosis and wrote the first draft of the manuscript and discussed the results. Author AMA designed the study. Author ABS performed the statistical analysis data. Author DQZ managed the analyses of the study. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v9i330225
Editor(s):
(1) Dr. R. Gayathri, Anna University, India.
Reviewers:
(1) Rao Muhammad Mahtab Mahboob, University of Agriculture,Faisalabad,Pakistan.
(2) A.Haja Alaudeen, Periyar Maniammai University, India.
(3) Praveena Nuthakki, V. R. Siddhartha Engineering College, India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/69216

Received 01 April 2021
Accepted 10 June 2021
Published 15 June 2021

Original Research Article

ABSTRACT

Leukemia refers to a disease that affects the white blood cells (WBC) in the bone marrow and/or blood. Blood cell disorders are often detected in advanced stages as the number of cancer cells is much higher than the number of normal blood cells. Identifying malignant cells is critical for diagnosing leukemia and determining its progression. This paper used machine learning with classifiers to detect leukemia types as a result, it can save both patients and physicians time and money. The primary objective of this paper is to determine the most effective methods for leukemia detection. The WEKA application was used to evaluate and analyze five classifiers (J48, KNN, SVM, Random Forest, and Naïve Bayes classifiers). The results were respectively as follows: 83.33%, 87.5%, 95.83%, 88.88%, and 98.61%, with the Naïve Bayes classifier achieving the highest accuracy; however, accuracy varies according to the shape and size of the sample and the algorithm used to classify the leukemia types.

**Corresponding author: E-mail: revella.eshaya@dpu.edu.krd;*

Keywords: Leukemia; WBC; machine learning; feature selection; WEKA application; J48; KNN; SVM; random forest; naïve bayes.

1. INTRODUCTION

Nowadays, machine learning (ML) is used in virtually every area of computational work that requires the design of algorithms and performance optimization. Learning from unbalanced data sets has become a critical issue in machine learning in recent years, and it is frequently encountered in a variety of applications, including computer security, engineering, remote sensing, biomedicine, and transformational industries [1,2].

Classification, regression, and band techniques are all examples of supervised learning techniques where the target variable is categorical and continues to decline. Machine learning datasets typically consist of a large number of samples and a limited number of attributes. Microarray technology is distinct from more conventional machine learning datasets [3].

Leukemia is a type of cancer of the blood cells. It is a disease that affects the delicate inner lining of the body called bone marrow. The bone marrow contains hematopoietic stem cells. It differentiates into a variety of blood components, including white blood cells (WBCs), platelets, and red blood cells (RBCs), each of which performs a unique function [4]. The (RBC) are responsible for transporting oxygen from the lungs to the body's tissues. While (WBC) are responsible for fighting disease and inflammation, platelets aid in clotting and bleeding control [5].

Different approaches for diagnosing leukemia have been developed using machine learning. One significant flaw remains in automated hematology analyses, and these studies sought to replicate the known parameters of those variables accurately. Computer algorithms are developed for aspects of imaging laboratory research such as hematological parameter analysis, blood value analysis, and other studies on hematological Analysis [6]. Diversification of classifiers is a critical aspect that plays a significant role in the application of this ensemble mechanism. This can be accomplished by utilizing various subsamples of the input data, as in the case of improvement [7], of course, there are some critical and fundamental differences between the categories of ensemble methods [8].

Additionally, in an ensemble, classifiers can be sequential or concurrent. The ensemble is classified into two types based on the learning mechanism used during the training phase and the testing of a subsample of the dataset: ensemble base learning and meta-learning [9]. Waikato Environment for Knowledge Analysis (WEKA) was used in this study to perform data mining techniques. WEKA is a piece of software that enables the execution of various data mining processes through the use of machine learning algorithms.

The feature selection procedure is frequently used to improve the effectiveness of all data mining algorithms and their performance in data classification [10]. The dataset contains numerous features, but not all of them are required. Certain features are redundant or irrelevant, providing no additional information and providing no context-relevant information, respectively. The feature selection is guided by a predefined criterion for selecting a subset of the original features and employs dimension reduction techniques frequently used in data mining [11].

Additionally, the feature selection process is used to reduce the number of features by eliminating those that are redundant, irrelevant, or noisy. This is especially advantageous because irrelevant features increase the complexity of the model and the time required to reach a stable model structure. Additionally, it is believed that the feature selection process speeds up the learning or modeling process, increases the accuracy or quality of the learning, and results in a better understanding of the model [12].

The rest of the paper is organized as follows. The second section mentions Related Work. Section 3 Leukemia Disease and its types. Section 4 Leukemia Dataset, Methodology in Section 5 describes the Feature Selection, classification used in detecting leukemia and Weka Tool. In section 6 comes performance evaluation matrices for classifier, also mentions the confusion matrix. The 7th section shown the experimental results and discussion. in section 8 comparative studies, the conclusion comes in the final section of the research.

2. RELATED WORK

S. Dasariraju et. al., [13] Extracted 16 features, two of which are novel features of the nucleus's color. A random forest algorithm was trained to detect and classify immature leukocytes. The model detected immature leukocytes with an accuracy of 92.99 percent and classified them with a classification accuracy of 93.45 percent. Precision values for each class were greater than 65 percent, which represents an improvement over the current state. The nucleus to cytoplasm area ratio was discovered to be a discriminative feature for both detection and classification, while the other two proposed features were shown to be significant for classification. The proposed model can aid in the diagnosis of AML, and the most salient features serve as a baseline for future research.

P. M. Gumble [14] propose the identification of leukemic blood cells through morphological analysis of microscopic images; morphological analysis requires only an image rather than a blood sample, making it ideal for low-cost and remote diagnostic systems. The proposed system first distinguishes leucocytes from other blood cells in the blood image, then selects lymphocyte cells (those associated with acute leukemia), evaluates their morphological indexes, and finally classifies the presence of leukemia. For each blood cell, the segmentation process generates two enhanced images, one of which contains the cytoplasm and the other of which contains the nuclei. The two images can then be used to extract distinctive features specific to each type of leukemia for identification. A total of 72 samples were collected from the 66 correctly identified samples using the KNN classifier to classify ALL. The system is 91.66 percent accurate.

U. K. Dey et. al., [15] purpose to use three machine learning algorithms to analyze the gene expression data of several individuals and predict the type of leukemia they have. XGBoost, Random Forest Classification, and Artificial Neural Networks are three of these algorithms. Prior to applying the algorithms, the dimensionality of the dataset was reduced via principal component analysis (PCA). The dataset was obtained through Kaggle website and contains the genetic expression profiles of 72 individuals, each of whom possessed 7129 genes. Meanwhile, the Precision value indicates how accurately the model predicts that an individual is afflicted with ALL. For Random Forest, this value is 73.7 percent. The accuracy remains constant from both perspectives for obvious reasons.

P. K. Mallick et. al., [16] pieced to classify gene expressions. There are 72 bone marrow expression datasets in the work that were used for the research classifier is for separating the acute-lymphocyte (ALL) and diffuse lymphocyte (AML) the network training uses 80% of the data, while the other 20% serves to validate the model. In comparison, it provided a good result. Although two types of leukemia have a 98.2% probability of being diagnosed, these probabilities have improved over time. These different computer-aided analyses may be of use to genetic and viral researchers in the future.

S. Mandal et. al., [17] created an image-based approach for cancer diagnosis by extracting critical information from the blood image data and training multiple classifiers. Others also claimed that Gradient Boosting Decision Tree (GBDT) classifiers give better results than Support Vector Machine (SVM) learning algorithms. the research also concluded that they had deduced several important characteristics, such as the presence of neighboring nuclei and the shape of the nucleus, which in turn influences the outcome of cell detection. This work stored vector graphics technique can be applied in a restricted computing environment without a Graphics Processing Unit. The program achieved an 85.6% of F1 validation score on data. has also located an important feature for helping physicians or technicians to quickly interpret stained images aid in the detection of leukemia patients.

A. Belhekar et. al., [18] proposed an image-analytics system that was completely automated. Using image analytics and classification algorithms on samples of patient's cells, the proposed system produced the correct results. TCIA (the Cancer Imaging Archive) has been used to collect the dataset for experiments. It has been prepared for processing. an “open source” predictive tool is implemented as “Orange-Mining”, KNN has shown itself to perform well for the segmentation of the data, while Neural Networks have shown to be superior for classification. the model's 0.865 AUC, 0.38 calculation precision, and F1 rating for neural networks.

T. Sajana et. al., [19] Presented 152 patients data, with Random split function for leukemia classification are analyzed clinically and presented in the paper. The classifier performance was also compared with several ensemble techniques – multiclass, Logit Boost, Stacking, and Random Committee classifiers. Experiments are conducted and confirmed that Random Bagging on the leukemia class dataset produces an accuracy of 95%.

P. K. Das et. al., [20] developed an algorithmic leukemia classification method to differentiate between various types of leukemia, an automated leukemia detection and classification procedure (ALL). Used datasets of 108 and 260 images respectively for ALL-IDB1, and ALL-IDB2. The cells were retrieved by processing with the color k-means clustering method that used a tool that extracted the lymphocytes. Finally, the segmentation process was completed, after which three main features were extracted: design, texture, and color. In some instances, other algorithms, such as the gray-level co-occurrence matrix (GLCM) and gray-level run-length matrix (GLRL) were then applied to enhance the features of the nucleus. Further, dimensional reduction with Principal Component Analysis (PCA) was used. WBCs were finally handled using an SVM (support vector machine) with an RBF kernel. 96% of the proposed method worked accuracy well, and sensitivity 92.64% of that efficacy was recorded.

E. Purwanti et. al., [21] propose an automated method for detecting lymphocyte leukemia by classifying single lymphocyte images obtained from peripheral blood smears. This study has two primary objectives. The first step is to isolate cells of interest. The second objective is to divide lymphocytes into two types, normal and abnormal lymphocytes. The authors combined shape and histogram features and used the k-nearest Neighbor algorithm with k values of 1, 3, 5, 7, 9, 11, 13, and 15. and the result was 90%, which was obtained by combining the characteristics of area-perimeter-mean-standard deviation with k=7.

S. Kumar et. al., [22] presented an algorithm for developing automated systems for detecting acute leukemia. The implemented method makes use of basic enhancement, morphology, filtering, and segmentation techniques to extract regions of interest using the k – means clustering algorithm. The proposed algorithm achieved a 92.8 % accuracy when compared to the K-Nearest Neighbor (KNN) and Naive Bayes Classifier on a 60-sample dataset.

3. LEUKEMIA DISEASE

Leukemia is a form of blood tissue cancer. The delicate inside of the body, called bone marrow, is leukemia. Leukemia. Hematopoietic stem cells are composed of the bone marrow. It evolves into multiple blood components such as white blood cells (WBCs), platelets, and red blood cells (RBCs), which each have distinct functions [4]. Cells (RBC) are responsible for the transfer of oxygen from the lungs to the tissues of the body. Though (WBC), also known as leukocytes, is responsible for combating disease and inflammation, platelets help with clotting and control bleeding [5]. The first type of leukemia consists of two categories: chronic leukemia and acute leukemia [23]. Acute leukemia is referred to as acute myeloid leukemia, acute lymphocytic leukemia, categorized as chronic myeloid leukemia, and chronic lymphocytic leukemia [24].

Leukemia is a serious disease in American culture, affecting all children and adults, as well as infants younger than 12 months. The most common type of cancer in children is leukemia, although the World Health Organization's study of adults indicates that leukemia is one of the 15 most common types of cancer [25]. A critical characteristic of cancer killers is the rapid development of irregular cells that grow beyond their natural boundaries, then invade neighboring cells and spread to other organs. This process is referred to as metastasis. According to the World Health Organization's website, cancer is the world's second leading cause of death, accounting for nearly 9.6 million deaths in 2018 [26]. According to a National Cancer Institute (NCI) report, the United States is expected to see 62,130 new cases of cancer treatment and 245,000 fatal or extremely serious cases [27].

3.1 Acute Myeloid Leukemia (AML)

Acute leukemia is the most common type. It occurs when the bone marrow begins to produce immature WBC and blisters. This also allows for the formation of irregular platelets and RBC. The symptoms may resemble those of influenza or other common diseases. Additionally, the terrain and markings may vary depending on the affected cell types. Acute myelogenous leukemia is typically characterized by fever, fatigue, and exhaustion, bone pain, phaleness of the skin, shortness of breath, and sudden swelling, as well as recurrent diseases and bleeding, such as hemorrhage gums and common nose bleeds. (AML) has eight additional subtypes that differ from other types of leukemia [28].

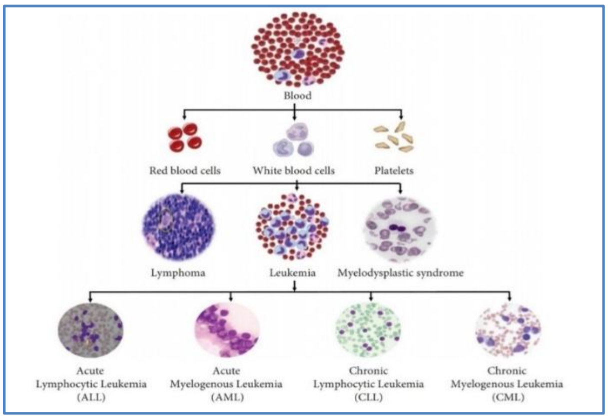

Blood
Red blood cells
White blood cells
Platelets
Lymphoma
Leukemia
Myelodysplastic syndrome
Acute Lymphocytic Leukemia (ALL)
Acute Myelogenous Leukemia (AML)
Chronic Lymphocytic Leukemia (CLL)
Chronic Myelogenous Leukemia (CML)

Fig. 1. Types of leukemia [28]

3.2 Acute Lymphocytic Leukemia (ALL)

It's the most common cancer in children, and it's caused by an overabundance of premature white blood cells in the bone marrow [29], as well as a long-term overabundance. It's difficult to tell the difference between flu and other common illnesses because they share symptoms like bone and joint fatigue, stiffness, and discomfort [5].

3.3 Chronic Myeloid Leukemia (CML)

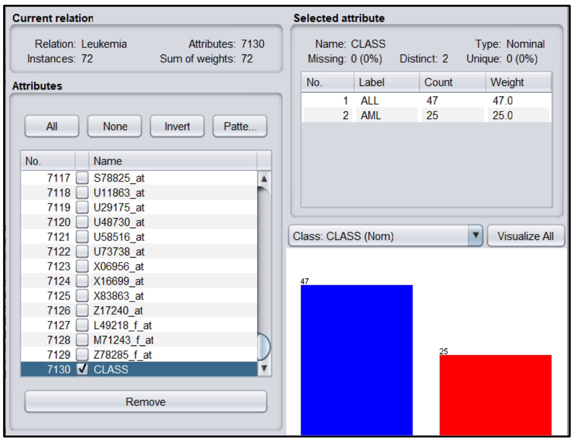
The Chronic myeloid leukemia is rare at a young age [30]. It is a slow-growing type of leukemia, and it can progress to fast-growing acute leukemia and difficult treatment. It can be seen in three phases, i.e., accelerated, chronic, and eruption phases. As it is in the chronic phase, leukemia grows slowly and is in the strongest case. The second stage, however, goes through a stage in which the blood cells are immature, usually known as the extended stage. Finally, it passes through the third stage, which is the explosion stage, known as the transformation stage of the explosion or the acute stage [28].

3.4 Chronic Lymphoblastic Leukemia (CLL)

It is known as a blood disease that slowly gets worse. It is not very common in children but is most commonly observed in adults, its symptoms include night sweats, fever, weight loss, and periodic infections [28].The types of leukemia and a pictorial representation of the blood structure are shown in Fig. 1.

4. LEUKEMIA DATASET

Leukemia dataset is microarray datasets in Weka ARFF format [31]. The database consists of 7130 features (7129 features (Genes) and 1 class attribute) for 72 instances (47 for ALL class, 25 for AML class), all are numerical values except the Last column is the class (two classes: ALL, AML) [32]. Fig. (2) shows the Weka information for Leukemia dataset.



Current relation

Relation: Leukemia Attributes: 7130
Instances: 72 Sum of weights: 72

Attributes

No.		Name
7117	□	S78825_at
7118	□	U11863_at
7119	□	U29175_at
7120	□	U48730_at
7121	□	U58516_at
7122	□	U73738_at
7123	□	X06956_at
7124	□	X16699_at
7125	□	X83863_at
7126	□	Z17240_at
7127	□	L49218_f_at
7128	□	M71243_f_at
7129	□	Z78285_f_at
7130	☑	CLASS

Selected attribute

Name: CLASS Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	ALL	47	47.0
2	AML	25	25.0

Class: CLASS (Nom)

Fig. 2. Leukemia dataset weka information

5. METHODOLOGY

Machine learning is the ability of computers to learn, where a machine is built with algorithms that allow it to make its own decisions and show the results to the user. It is essentially known as the subfield of Artificial Intelligence. Machine Learning is used today to classify and make decisions on complex data. In general, algorithm development enables the machine to learn and make the necessary decisions. It is closely related to mathematical optimization, which supplies the field with tools, theory, and implementation domains, and is used in a variety of computational tasks where explicit algorithms cannot be planned and programmed. Machine Learning techniques and tasks are broadly classified into three categories:

- Supervised learning: which is capable of solving regression problems such as weather forecasting, population growth forecasting, and life experience forecasting, among others, is accomplished through supervised learning using either a Linear Regression or a Random Forest algorithm. Additionally, supervised learning solves classification problems such as voice recognition, digit recognition, diagnostics, and identity fraud detection by utilizing algorithms from a variety of fields, including Support Vector Machines, Random Forests, and K-Nearest Neighbor. There are two levels of supervised learning. The phase of training and the phase of testing. The data sets used in the training process must contain known labels. The algorithms examine the relationship between the input values and the labels and make predictions about the values of the testing data [33].

- Unsupervised Machine Learning (UML), a widely used technique for analyzing multi-omics data, has the potential to significantly advance our understanding of patient phenotypes and clinical outcomes [34]. On the other hand, clinical data is more heterogeneous than high-throughput datasets, posing unique challenges for UML. While handling mixed data is critical in bioinformatics, UML is frequently used in omics contexts to apply a single distance metric uniformly to a matrix of homogeneous data, either continuous or binary [35].

- Predicting patient outcomes using machine learning techniques has demonstrated superior accuracy to other methods. This is why machine learning has been a hot topic of research in recent years. For example, machine learning techniques have been used to forecast the outcome of various types of cancer [36], and Numerous classification techniques have been developed in the field of machine learning, and a significant number of them have been applied to the classification of cancer [37].

In this proposed model many classifiers were used to classify the Leukemia with high accuracy, efficiency using J48, KNN, SVM, Random Forest, and Naïve Bayes classifiers. The mechanism of this proposed model goes through four main stages, which are (1-Leukemia dataset uploading, 2-feature selection, 3-classification and 4-evaluating the results). Fig. (3) shows the flowchart diagram of the proposed model. After the classification results of the five algorithms, the performance was measured by calculating the following metrics: True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, ROC Area.

As shown in Fig. 3, 10-fold cross-validation was utilized. Fold Cross-validation is a method for evaluating predictive models that splits the dataset into a training subset and a test subset for training and evaluating the model. The Leukemia dataset was randomly partitioned into 10 equal size subgroups in 10-fold cross-validation. One subset is kept for validation, while the other nine are used for training purposes. The folds are then used to repeat cross-validation ten times, with each of the ten subsets serving as validation data exactly once. The results of the ten folds can then be averaged to get a single estimate. This method has the advantage of using all observations for training and validation, and each observation is only utilized once for validation.

5.1 Feature Selection

The Weka platform includes feature selection algorithms that use filter approaches to choose relevant parameters and improve the performance of machine learning models [38]. In this paper for the feature selection CorrelationAttributeEva (CA) method were used. CA is a feature subset selection algorithm [39]. It evaluates the attribute by calculating the correlation (Pearson's product moment correlation) between it and the class [40,41]. The main objective of CA is to obtain a highly relevant subset of features that are uncorrelated to each other.

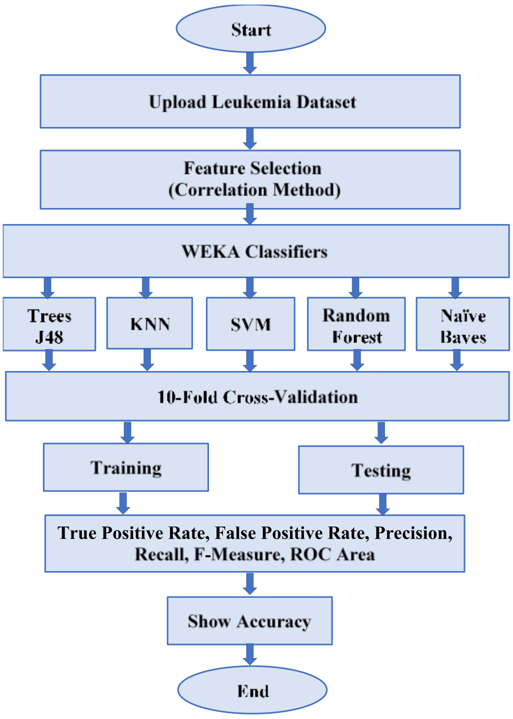


Fig. 3. Leukemia diagnosis proposed model flowchart diagram

In this way, the dimensionality of datasets can be drastically reduced and the performance of learning algorithms can be improved [42]. Ranker search method used with CA. Features are prioritized and those that are most suited for use in the machine learning method are filtered based on their Correlation values [38,43]. By the combination of this Correlation attribute evaluator with the Ranking method of Search is applied to the Leukemia dataset.

5.2 Classification

Classification is the theoretical process of identifying individual components of an image or even the entire image. Image categorization is a critical and widely used branch of image processing that involves classifying images into a set of predefined categories using only representative samples from each category [44,45]. There are two types of image classification; an uncontrolled image classification exists [46]. Although the current work focuses on mild classification algorithms, they are not unexpected. The primary classification in machine vision is also regulated [14].

5.2.1 Support vector machines (SVM)

The SVM classifier is based on the hyper plane classifier. This is calculated through a feature (linear or non-linear) It serves as a classifier [47]. features extracted and captured from the image texture and shape One of the important features of the WBC class is the length and number of the lobes. relevant features for training, like the number of lobes, periphery/nuclei, nuclei count, nuclei/entropy, and a number of cells per species are selected [48].

5.2.2 K-nearest neighbor (KNN)

KNN classifies new states based on a similarity measure (for example, Euclidean distance functions) This method was already used in the 1970's for pattern recognition and statistical estimation [14]. Additionally, a method of training classifier optimization called K-NN was used to search for the k-species in the training datasets while incorporating data variability [49].

5.2.3 Decision tree (J48)

J48 is a more developed version of C4.5 that is designed to work with continuous data [12]. The method initially constructs a tree using the training data via a training stage. A sample from the testing data is compared to the constructed tree in order to determine its class. In fact, Decision Trees are widely used by numerous scholars and are considered to be one of the simplest classifiers to use. It is designed and developed based on data entropy [50]. Due to the technique's tree-like shape and form, it is one of the most accurate and time-efficient classifiers. The anxious subdivision is regarded as a symbol for the conclusion of conceivable feature standards through without a doubt [51].

5.2.4 Random forest (RF)

This technique constructs a forest by combining multiple decision trees in order to achieve a high classification rate [52]. The ultimate goal of utilizing this supervised classifier machine is to avoid over-reliance on a single learning model [25]. The critical distinction between this novel technique and the conventional decision tree classifier is that the root nodes contain divided nodes that are not necessary [53].

5.2.5 Naïve bayes

The goal of the Naïve Bayes algorithm is to detect blast cells based on features They utilize the proposed classifier to detect the presence of blast cells. Combining the Naïve Bayes classifier with a gene classification may help The job of the classifier is to collect the Naïve Bayes results It is clear and effective. Convergence causes social inequity and takes measures [54].

5.3 Weka Tool

Given that the output of the training dataset is known to be the screening class, classification is the most appropriate technique to use in this case. Classification can aid in the improvement of screening procedures and the reduction of potential errors caused by inexperienced health professionals.

WEKA is a set of machine learning algorithms for data mining. The algorithms can be applied to a dataset directly or via Java code. WEKA has tools for data preprocessing, classification, regression, clustering, and visualization. It is also well-suited to the development of new machine learning systems. WEKA is open-source software released under the GNU General Public License. Additionally, it is a self-contained platform, as the program is written in the Java ™ programming language and includes a graphical user interface for interacting with data files and generating visual results (tables and curves thinking). Additionally, it includes a generic API, which enables you to automatically include WEKA, like any other library, in our applications for tasks such as server-side data mining [55]. Fig. (2) shows the block diagram of the proposed model.

6. PERFORMANCE EVALUATION MATRICES FOR CLASSIFIER

To evaluate the best classifier, we must first define some well-known performance metrics that will assist us in selecting the best.

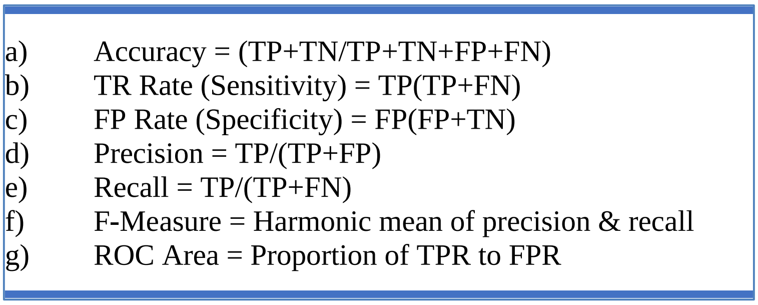
The Confusion Matrix provides an excellent overview of a classifier's performance. A typical confusion matrix is shown in Table 1.

Table 1. A Typical 2*2 confusion matrix

Actual Class	**Predicate Class**	
	Positive	**Negative**
Positive	**TP**	**FN**
Negative	**FP**	**TN**

TP = No. of Positive Classified Correctly Instance.
FN = No. of Negative Classified Incorrectly Instance.
FP = No. of Positive Classified Incorrectly Instance.
TN = No. of Negative Classified Correctly Instance.

Confusion Matrix gives a Number of performance measure



a) Accuracy = (TP+TN/TP+TN+FP+FN)
b) TR Rate (Sensitivity) = TP(TP+FN)
c) FP Rate (Specificity) = FP(FP+TN)
d) Precision = TP/(TP+FP)
e) Recall = TP/(TP+FN)
f) F-Measure = Harmonic mean of precision & recall
g) ROC Area = Proportion of TPR to FPR

7. EXPERIMENTAL RESULTS AND DISCUSSION

This section summarizes the results of the classification process in which we applied the classifiers to the dataset using WEKA Experimenter.

Table 2 shows and analysis the results for using J48 algorithm.

Table 3 shows and analysis the results for using KNN algorithm.

Table 4. shows and analysis the results for using SVM algorithm.

Table 5 shows and analysis the results for using Random Forest algorithm.

Table 6 shows and analysis the results for using Naïve Bayes algorithm

Table 7 shows the comparison between the five classifiers depending on the time taken to build the model and the accuracy of the classifier.

Fig. 4 shows that random forest algorithm tokes the longest time to build its model while KNN algorithm had the shortest time.

Fig. 5 shows that Naïve Bayes algorithm had the higher accuracy.

Table 2. System assessment using trees J48 classifier

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
ALL	0.894	0.280	0.857	0.894	0.875	0.784
AML	0.720	0.106	0.783	0.720	0.750	0.784
Weighted Avg.	0.833	0.220	0.831	0.833	0.832	0.784

Table 3. System assessment using KNN classifier

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
ALL	0.936	0.240	0.880	0.936	0.907	0.875
AML	0.760	0.064	0.864	0.760	0.809	0.875
Weighted Avg.	0.875	0.179	0.874	0.875	0.873	0.875

Table 4. System assessment using SVM classifier

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
ALL	0.979	0.080	0.958	0.979	0.968	0.949
AML	0.920	0.021	0.958	0.920	0.939	0.949
Weighted Avg.	0.958	0.060	0.958	0.958	0.958	0.949

Table 5. System assessment using random forest classifier

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
ALL	1.000	0.040	0.979	1.000	0.989	0.980
AML	0.960	0.000	1.000	0.960	0.980	0.980
Weighted Avg.	0.986	0.026	0.986	0.986	0.986	0.980

Table 6. System assessment using naïve bayes classifier

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
ALL	1.000	0.320	0.855	1.000	0.922	0.986
AML	0.680	0.000	1.000	0.680	0.810	0.986
Weighted Avg.	0.889	0.209	0.905	0.899	0.883	0.986

Table 7. Comparison of results by time and accuracy

Classifier	Comparison of results	
	Time taken to build model	Accuracy
Trees J48	0.22 Sec.	83.3333 %
KNN	0.05 Sec.	87.5 %
SVM	0.21 Sec.	95.8333 %
Random Forest	1.4 Sec.	88.8889 %
Naïve Bayes	0.22 Sec.	98.6111 %

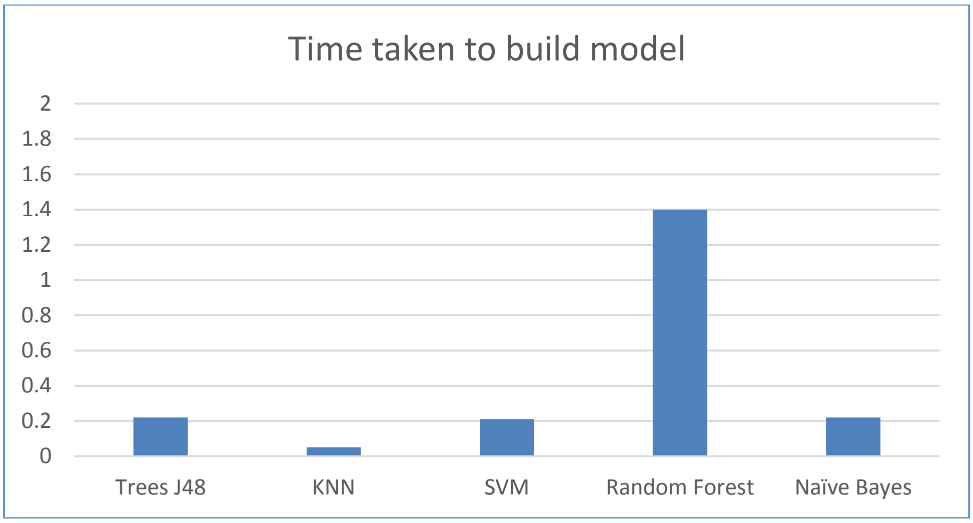


Fig. 4. Time analysis

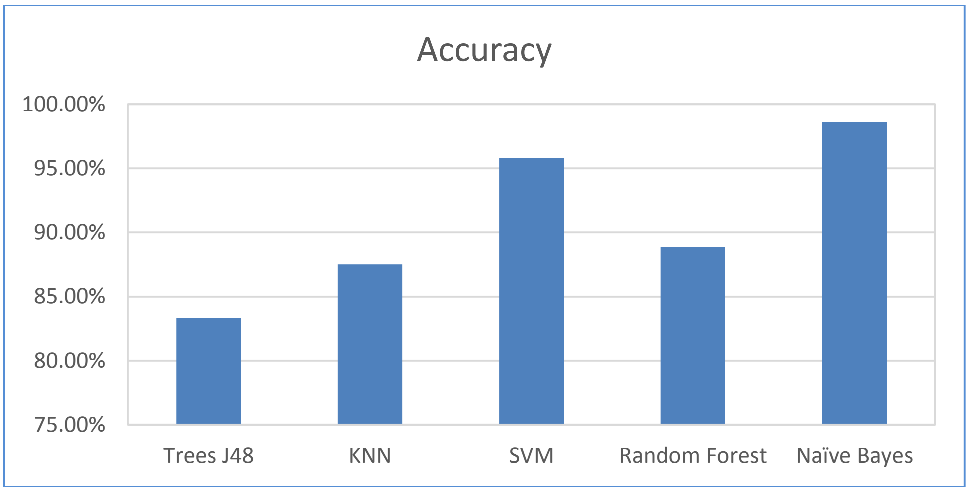


Fig. 5. Accuracy analysis

8. COMPARATIVE STUDIES

Table 8 shows a summary of comparison for related work. This table demonstrates that researchers in related papers used a variety of different methods for trait selection and classification, as well as a variety of different datasets with varying numbers of leukemia samples.

Although dataset was different, previous research papers mostly used the same classifiers to extract accurate results for leukemia. Whereas, we notice that, the research paper [16] showed higher accuracy results with a value of 98.2% by using classifier for separating on microarray Dataset. Whereas for [15] the result was 80.8% by using the Random Forest classifier with PCA Feature Selection on ALL microarray Dataset. Additionally, using KNN and Naive Bayes in [22] the result was 92.8% with KNN classifier on Clinically Collected Dataset. While the [20] used KNN and SVM classifiers and showed 91.20%, 96.00% accuracy in the same dataset with (GLCM) & (GLRLM) feature selection.

To evaluate the accuracy of each classifier, a confusion matrix is used. The experimental result demonstrates that when five attributes are applied, Naïve Bayes archives the optimum recognition ratio with 98.61% and second is the SVM with 95.83%, while KNN accuracy ratio is 87.5%. Meanwhile, J48 gets 83.33% of recognition ratio, and Random Forest gets 88.88% of recognition ratio. The based on the accuracy results, we can see that the best results appear when the ROC Area value is close to 1.

Table 8. Comparison this study and related work

Reference /Year		Comparison			
		Dataset	**Feature Selection**	**Classifier**	**Result**
[12]	2020	Clinically Collected Dataset	-	Random Forest	92.99% 93.45%
[15]	2020	Microarray ALL	-	Classifier for Separating	98.2%
[18]	2020	Clinically Collected Dataset	-	Multiclass, Logit Boost, Random Committee, Stacking, Random Split Committee	75% 89% 90% 60% 95%
[19]	2020	Clinically Collected Dataset	(GLCM) & (GLRLM)	KNN SVM	91.20% 96.00%
[14]	2019	Microarray ALL	PCA	Xgboost & Random Forest	92.3% 80.8%
[16]	2019	TCIA	Lightgbm Model	SVM & GBDT Classifier	79.4 85.6%
[17]	2019	TCIA	Orang Mining	Random Forest KNN	82.5% 82.7%
[21]	2018	Clinically Collected Dataset	KNN	KNN Naive Bayes	92.8%
[13]	2017	Clinically Collected Dataset	-	KNN	91.66%
[20]	2017	Clinically Collected Dataset	-	KNN	90%
Proposed Work		Microarray ALL	Correlation	Trees J48 KNN SVM Random Forest Naïve Bayes	83.3333% 87.5% 95.8333% 88.8889% 98.6111%

9. CONCLUSION

The primary objective of this paper is improving the accuracy of leukemia using machine learning with classification Algorithms. Microarray Leukemia dataset were used and classified using five classifiers (J48, KNN, SVM, Random Forest and Naïve Bayes algorithms). The results show that Naïve Bayes classifier has the highest accuracy 98.61%, where J48 has the lowest accuracy 83.33%. The last three classifiers respectively are KNN, SVM, Random Forest with the accuracy ratios of 87.5%, 95.83%, 88.88%, and 95.4%. Further work in this research will lead to determining the effectiveness of treatment provided to leukemia patients, through effective use of appropriate machine learning classification algorithms of all types of leukemia, which can be executed in parallel for better response time and accuracy.

DISCLAIMER

The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Abdulqader DM, Abdulazeez AM, Zeebaree DQ. Machine learning supervised algorithms of gene selection: A review. Technol. Reports Kansai Univ. 2020;62(3):233–244.
2. Sengar PP, Gaikwad MJ, Nagdive AS. Comparative study of machine learning algorithms for breast cancer prediction.

Proc. 3rd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2020, no. September. 2020;796–801.
DOI: 10.1109/ICSSIT48917.2020.9214267

3. Zeebaree DQ, Haron H, Abdulazeez AM. Gene Selection and classification of microarray data using convolutional neural network. ICOASE 2018 - Int. Conf. Adv. Sci. Eng. 2018;145–150.
DOI: 10.1109/ICOASE.2018.8548836
4. Kumar S, Mishra S, Asthana P. Automated detection of acute leukemia using k-mean clustering algorithm; 2018.
5. Shafique S, Tehsin S. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. 2018;17:1–7.
DOI: 10.1177/1533033818802789
6. Jagadev P. Using image processing and machine learning. 2017;522–526.
7. Triayudi A. Alg clustering to analyze the behavioural," no; 2018.
8. Triayudi A. "Data mining implementation to predict sales using time series method," no; 2020.
DOI: 10.11591/eecsi.v7.2028
9. Fabrianne SF, Triayudi A, Sholihati ID. Data mining using filtering approaches and ensemble methods; 2021.
DOI: 10.1088/1757-899X/1088/1/012012
10. Armya REA. Medical images segmentation based on unsupervised algorithms : A review.
DOI: 10.48161/Issn.2709-8206
11. Eesa AS, Orman Z, Brifcani AMA. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Syst. Appl. 2015;42(5):2670–2679.
DOI: 10.1016/j.eswa.2014.11.009
12. Aljawarneh S, Bani M, Mohammed Y. An enhanced J48 classification algorithm for the anomaly intrusion detection systems. Cluster Comput; 2017.
DOI: 10.1007/s10586-017-1109-8
13. Dasariraju S, Huo M, McCalla S. Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm. Bioengineering. 2020;7(4):1–12.
DOI: 10.3390/bioengineering7040120
14. Gumble PM. Analysis and classification of acute lymphoblastic leukemia using KNN algorithm. 2017;94–98.
15. Dey UK, Islam MS. Genetic expression analysis to detect type of leukemia using machine learning. 1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019, vol. 2019, no. Icasert. 2019;1–6.
DOI: 10.1109/ICASERT.2019.8934628
16. Mallick PK, Mohapatra SK, Chae GS, Mohanty MN. Convergent learning–based model for leukemia classification from gene expression. Pers. Ubiquitous Comput; 2020.
DOI: 10.1007/s00779-020-01467-3
17. Mandal S, Daivajna V, Rajagopalan V. Machine learning based system for automatic detection of leukemia cancer cell. 2019 IEEE 16th India Counc. Int. Conf. INDICON 2019 - Symp. Proc. 2019;1–4.
DOI: 10.1109/INDICON47234.2019.9029034
18. Belhekar A, Bhelkar Y, Gagare K, Rajeswari K, Bedse R, Karthikeyan M. Leukemia cancer detection using image; 2019.
19. Sajana T, Maguluri LP, Syamala M, Usha Kumari C. Classification of leukemia patients with different clinical presentation of blood cells. Mater. Today Proc., no. xxxx; 2020.
DOI: 10.1016/j.matpr.2020.10.619
20. Das PK, Jadoun P, Meher S. Detection and classification of acute lymphocytic leukemia. Proc. 2020 IEEE-HYDCON Int. Conf. Eng. 4th Ind. Revolution, HYDCON 2020; 2020.
DOI: 10.1109/HYDCON48903.2020.9242745
21. Purwanti E, Calista E. Detection of acute lymphocyte leukemia using k-nearest neighbor algorithm based on shape and histogram features. J. Phys. Conf. Ser. 2017;853(1).
DOI: 10.1088/1742-6596/853/1/012011
22. Kumar S, Mishra S, Asthana P, Pragya. Automated detection of acute leukemia. Adv. Intell. Syst. Comput. 2018;554:655–670.
DOI: 10.1007/978-981-10-3773-3_64
23. Thanh TTP, Vununu C, Atoev S, Lee S, Kwon K. Leukemia blood cell image classification using convolutional neural network. 2018;10(2).
24. HA. Segmentation of leukemia cells using clustering. 2019;10(2):39–48.
DOI: 10.4018/IJSE.2019070103
25. Sachin P, Kumar RY. Detection and classification of blood cancer from microscopic cell Images using SVM KNN

and NN classifier. Int. J. Adv. Res. 2017;3(6):315–324, , [Online]. Available: www.ijariit.com.

26. K-means HSCR. Leukemia image segmentation using a hybrid clustering algorithm. 2020;1–22,.
27. Najat N, Abdulazeez AM. Gene clustering with partition around mediods algorithm based on weighted and normalized mahalanobis distance. ICIIBMS 2017 - 2nd Int. Conf. Intell. Informatics Biomed. Sci. 2018;140–145. DOI: 10.1109/ICIIBMS.2017.8279707
28. Bibi N, Sikandar M, Din IU, Almogren A, Ali S. IoMT-based automated detection and classification of leukemia using deep learning; 2020.
29. Abas SM, Abdulazeez AM. Detection and classification of leukocytes in leukemia using YOLOv2 with CNN. 2021;8(3):64–75. DOI: 10.9734/AJRCOS/2021/v8i330204
30. Lymphoma C, No S, Erimbetova I. Abstracts from proceedings of the society of hematologic oncology 2020 annual meeting rare case of Chronic Myeloid Leukemia Secondary to Acute Lymphoblastic Leukemia in a Young Adult . A Case Report the Population of the Republ. 2020;20:159–172. DOI: 10.1016/S2152-2650(20)30481-X
31. Zhu Z, Ong YS, Zurada JM. Identification of full and partial class relevant genes. IEEE/ACM Trans. Comput. Biol. Bioinforma. 2010;7(2):263–277. DOI: 10.1109/TCBB.2008.105
32. Kaggle G. et Al. “Gene expression dataset,” [Online]. Available:https://www.kaggle.com/crawford/gene-expression
33. Abdullah DM, Ahmed NS. A review of most recent lung cancer detection techniques using machine learning. 2021; 159–173. DOI: 10.5281/zenodo.4536818
34. Salim NOM, Abdulazeez AM. Human diseases detection based on machine learning algorithms : A review. 2021;102–113. DOI: 10.5281/zenodo.4467510
35. Coombes CE, Abrams ZB, Li S, Abruzzo LV, Coombes KR. Unsupervised machine learning and prognostic factors of survival in chronic lymphocytic leukemia. 2020;27:1019–1027. DOI: 10.1093/jamia/ocaa060
36. Omobolaji R, Sc AM, Sc MED, Ricardo D. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer; 2019.
37. Alrefai N. Ensemble machine learning for leukemia cancer diagnosis based on microarray datasets. 2019;14(21):4077–4084.
38. Kumar S, Chong I. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. Int. J. Environ. Res. Public Health. 2018;15(12). DOI: 10.3390/ijerph15122907
39. Al Janabi KB, Kadhim R. (‘Weka’ Feature Selection - bad results) data reduction techniques: A comparative study for attribute selection methods. Int. J. Adv. Comput. Sci. Technol. 2018;8(1):1–13. [Online]. Available:http://www.ripublication.com
40. Sugianela Y, Ahmad T. Pearson correlation attribute evaluation-based feature selection for intrusion detection system. Proceeding - ICoSTA 2020 2020 Int. Conf. Smart Technol. Appl. Empower. Ind. IoT by Implement. Green Technol. Sustain. Dev. 2020;1–5. DOI: 10.1109/ICoSTA48221.2020.1570613717
41. GS, TM, MVT, GV. Classification algorithms with attribute selection: An evaluation study using WEKA. Int. J. Adv. Netw. Appl. 2018;9(6)L3640–3644.
42. Demisse GB, Tadesse T, Bayissa Y. Data mining attribute selection approach for drought modelling: A case study for greater horn of Africa. arXiv. 2017;7(4). DOI: 10.5121/ijdkp.2017.7401
43. Abdullah DM, Abdulazeez AM. “Lung cancer prediction and classification based on correlation selection method using machine learning techniques,” no. Ml. DOI: 10.48161/qaj.v1n2a58
44. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J. Appl. Sci. Technol. Trends. 2021;2(01):20–28. DOI: 10.38094/jastt20165
45. Khorshid SF, Abdulazeez AM, Sallow AB. A comparative analysis and predicting for breast cancer detection based on data mining models. Asian J. Res. Comput. Sci. 2021;8(4):45–59. DOI: 10.9734/ajrcos/2021/v8i430209

46. Zebari DA, Zeebaree DQ, Abdulazeez AM, Haron H, Hamed HNA. Improved threshold based and trainable fully automated segmentation for breast cancer boundary and pectoral muscle in mammogram images. IEEE Access. 2020;8:203097–203116.
DOI: 10.1109/access.2020.3036072.
47. Abdullah DM, Abdulazeez AM. Machine learning applications based on SVM classification : A review. 2021;81–90.
DOI: 10.48161/Issn.2709-8206
48. Engineering T. Acute leukemia classification by using SVM and K-Means clustering. 2014;1–4.
49. Wisesty UN, Warastri RS, Puspitasari SY. Detection of acute lymphocyte leukemia using k- nearest neighbor algorithm based on shape and histogram features Detection of acute lymphocyte leukemia using k-nearest neighbor algorithm based on shape and histogram features.
50. Maiwan BSRM, Mayyadah R. An analytical appraisal for supervised classifiers' performance on facial expression recognition based on relief-F feature selection; 2021.
DOI: 10.1088/1742-6596/1804/1/012055
51. Panigrahi R, Borah S. Science direct science direct rank allocation to J48 Group of decision tree classifiers using binary and multiclass intrusion detection datasets. Procedia Comput. Sci. 2018;132:323–332.
DOI: 10.1016/j.procs.2018.05.186
52. Abdulkareem NM, Abdulazeez AM. Machine learning classification based on radom forest algorithm : A review. 2021;128–142.
DOI: 10.5281/zenodo.4471118
53. Elyusufi Y, Elyusufi Z. "Social networks fake profiles detection using machine learning algorithms," no; 2020.
DOI: 10.1007/978-3-030-37629-1
54. Das BK, Dutta HS. "GFNB : Gini index – based Fuzzy Naive Bayes and blast cell segmentation for leukemia detection using multi-cell blood smear images"; 2020.
55. Ahmed W, Saeed A, Salah A, Abdala E. A comparative study on machine learning tools using WEKA and rapid miner with classifier algorithms C4.5 and decision stump for network intrusion detection". 2019;4(4):749–752.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/69216