DOI: 10.36108/ujees/5202.70.0141
UNIOSUN Journal of Engineering and Environmental Sciences. Vol. 7 No. 1. March. 2025

Classification of Leukemia Cancer Data using Correlation Based Feature Selection Model: A Comparative Approach

Babatunde, R. S., Isiaka, R. M., Abdulsalam, S. O., Arowolo, O. M., Ajao, J. F.

Abstract The abundance of data obtained from microarray experiments presents challenges related to the number of variables and the presence of random fluctuations. Despite the efforts that had been made by previous researchers, emphasizing how data mining aids the implementation of models to facilitate informed prediction, gaps are evident which requires improvement over the earlier models. Dimensionality reduction techniques, such as Correlation Based Feature Selection (CBFS), are good candidate solutions to these problems by selecting pertinent features for categorization. This research implements a model for classification of leukemia cancer using CBFS with Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree (DT), and Ensemble classifiers. The evaluation of the performance of these machine learning models was carried out using sensitivity, specificity, precision and accuracy. The findings indicate that the CBFS+DT model outperforms the other models in terms of sensitivity (96.75%), specificity (97.18%), precision (97.56%), accuracy (96.75%), and F1 score (96.97%), while also exhibiting a decreased computational time (0.4336). This demonstrates the efficacy of CBFS in improving classification accuracy and reducing computing load. Overall, this study highlights the effectiveness of CBFS in cancer research and underscores the importance of carefully choosing the most pertinent variables to enhance classification outcomes.

Keywords: machine learning, classification, feature selection, pattern recognition.

I. Introduction

A. Background to Study

Data mining (DM) has emerged as one of the most valuable methods for extracting and modifying data as well as for identifying patterns to generate information that can be used to make decisions in a variety of industries, including business, healthcare, and finance. In the healthcare sector, data mining contributes to the effectiveness of disease prevalence control, diagnosis, prediction, and prescription [1]. One of the main areas of study in the medical sector is cancer research, as it is important to accurately predict different types of tumors while providing better care for patients. Using gene expression data, the recent development of microarray technology has motivated the simultaneous monitoring of genes and cancer classification. The outcome obtained so far is promising in its early stage of growth [2], [3], [4]. The need for accurate cancer classification and prediction has led to the application of data mining and machine learning techniques in cancer research. The use of DNA microarray technology, which allows the analysis of gene expression data to categorize cancer phenotypes and forecast patient outcomes, is a result of machine learning techniques in cancer research. Major applications of DNA microarray technology are to perform sample classification analyses between different disease phenotypes, for diagnostic and prognostic purposes. The classification analyses involve a wide range of algorithms such as

Babatunde, R. S., Isiaka, R. M., Abdulsalam, S. O., Ajao, J. F.
(Department of Computer Science. Kwara State University, Malete. Nigeria)
Arowolo, O. M
(Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, United States)
Corresponding Author:: ronke.babatunde@kwasu.edu.ng

differential gene expression analyses, clustering analyses and supervised machine learning. Machine learning algorithms are most frequently used to complete this task [1].

Dimensionality reduction eliminates irrelevant features, reduce noise, and produce more robust learning models due to the involvement of fewer features. In general, the dimensionality reduction by selecting new features which are a subset of the old ones is known as feature selection [3]. Feature selection technique is a knowledge discovery tool which provides an understanding of the problem through the analysis of the most relevant features. Feature selection aims to select a subset of relevant features that are necessary and sufficient to describe the target concept. Feature selection is the process of reducing the dimensionality of the available data, with the aim of improving the recognition results [5]. This process typically consists of three steps: a search procedure for searching the solution space made of all the possible solutions, i.e., feature subsets, an evaluation function, and a stopping criterion. Filter methods measure statistical or geometrical properties of the subset to be evaluated, whereas wrapper functions adopt as evaluation measure accuracy achieved by a given, previously chosen, classifier [4]. Embedded approaches include feature selection in the training process, thus reducing the computational costs due to the classification process needed for each subset. Wrapper methods are computationally costly because the evaluation of each subset requires the training of the adopted classifier [6]. For this reason, they are typically used with near-optimal search strategies, which can achieve acceptable results, but limiting the computational costs. As for the filter methods, they need non-iterative computations on the dataset which are, in most of the cases, significantly faster than classifier training sessions [7]. Category of filter methods is that of the ranking ones, which evaluate each feature singularly. Once all features have been evaluated, they are ranked according to their merit. Then the subset search step is straightforward: the best M features are selected, with M set by the user. If this approach is very fast and allow dealing with thousands of features, there is no one general criterion for choosing the dimension of the feature space, then it is difficult to select the number M of features to be selected. Moreover, most importantly, relevant features that are highly informative when combined with other ones could be discarded because they are weakly correlated with the target class [7], [8]. CBFS is a simple multivariate filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function [10]. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. The CBFS technique is proposed in this research to select the most relevant features of leukemia cancer dataset in order to ensure a better classification model is built. Additionally, the model was implemented using SVM, KNN, DT, and ensemble learning methods and the performance was evaluated based on accuracy, precision, sensitivity, specificity and computation time.

B. Related Works

A deep learning framework was developed to detect leukemia in microscopic blood samples by [9] using Squeeze and Excitation Learning. The proposed deep learning architecture emphasizes channel associations on all feature representation levels by adding squeeze and excitation learning,

which recursively recalibrates channel-wise feature outputs by modeling channel interdependencies explicitly. The squeeze-and-excite technique improves the feature discriminability of leukemic and normal cells, exposes informative leukemia cell traits while suppressing less significant ones, and improves deep learning algorithm feature representational capacity. The work show that combining squeeze and excite in a deep learning model improves its ability to diagnose leukemia from blood samples. The proposed model was evaluated using ALL IDB1 and ALL IDB2 of leukemia patient blood samples, giving rise to a reliable computer-aided leukemia cancer diagnosis. A modified Firefly Deep Ensemble was proposed for microarray data classification in [10]. The work employed a Modified Firefly Feature Selection (MFFS) approach to remove irrelevant categorization features and a Deep Learning Model for microarray data classification. Experimental results show that the proposed MFFS algorithm combined with a Hybrid Deep Learning Algorithm outperforms existing approaches in feature set size, accuracy, precision, recall, F-measure, and AUC for a dataset with more features. A systematic review of leukemia detection and classification using smear blood images was conducted in [11]. Machine Learning (ML), leukemia, peripheral blood smear (PBS) picture, detection, diagnosis, and classification were used to search PubMed, Scopus, Web of Science, and ScienceDirect and Google Scholar. 116 items were found. 16 papers met the study's inclusion and exclusion criteria. The review examines all published ML-based leukemia detection and classification models that handle PBS pictures. The average accuracy of ML algorithms used to diagnose leukemia in PBS images was >97%, demonstrating that ML might lead to exceptional results. Deep learning (DL) had higher precision and sensitivity than other ML algorithms in diagnosing malignancies. The work concluded that ML improves accuracy, reduces diagnosis time, and provides faster, cheaper, and safer diagnostic services. Five supervised feature selection techniques for cancer multi-omics data were compared in [12]. mRMR, INMIFS, DFS, SVM-RFE-CBR, and VWMRmR for multi-omics datasets were used. Five feature selection algorithms are evaluated using three criteria: classification accuracy (Acc), representation entropy (RE), and redundancy rate (RR). Each feature subset's classification accuracy (Acc) was measured using C4.5, NaiveBayes, KNN, and AdaBoost. VWMRmR optimizes three datasets'Acc (ExpExon, hMethyl27 and Paradigm IPLs). The VWMRmR technique yields the best RR (obtained using normalized mutual information) for three datasets (Exp, Gistic2 and Paradigm IPLs) (Gistic2 and Paradigm IPLs). It's the best RE for three datasets (Exp, Gistic2 and Paradigm IPLs). VWMRmR performs best for all three evaluation criteria in most datasets. A Highly Discriminative Hybrid Feature Selection Algorithm for Cancer Diagnosis was implemented by [13]. Two-stage hybrid feature selection was used. In the first stage, an overall ranker combines chi-squared, F-statistic, and mutual information results (MI). This combination orders the features. Second, modified wrapper-based sequential forward selection was used to find the appropriate feature subset utilizing ML models such as SVM, DT, RF, and KNN classifiers. To test the technique, 10-fold cross-validation and hyperparameter adjustment were used on four malignant microarray datasets.. Both SVM and KNN models achieve 100% accuracy on the leukemia dataset using 5 characteristics. SVM model obtains 100% accuracy on ovarian cancer dataset using only 6 characteristics. SVM achieves 100% accuracy on the SRBCT dataset using only 8 features. For lung cancer, the SVM model achieves 99.57% accuracy utilizing 19 characteristics. In terms of selected features and

diagnostic accuracy, the suggested method outperforms other algorithms. [14] proposed a new statistical learning approach for ultrahigh-dimensional gene expression multi-classification. The work used model-free feature screening to retain informative gene expression values from ultrahigh-dimensional data, then build prediction models with gene expression network architectures. The outcome was a discovery of gene expression predictors and dependencies, unlike existing supervised learning methods. Analysis of a real dataset shows that the proposed approach provides precise classification and accurate prediction, outperforming other conventional supervised learning methods. [15] presents a hybrid cancer classification approach that uses several machine learning techniques: Pearson's correlation coefficient as a correlation-based feature selector and reducer, an easy-to-interpret Decision Tree classifier, and Grid Search CV (cross-validation) to optimize the maximum depth hyperparameter. The approach was evaluated using 7 microarray cancer datasets. To determine which model features are most useful and relevant, classification accuracy, specificity, sensitivity, F1-score, and AUC are used. The suggested technique reduces the number of genes needed for categorization, picks the most informative traits, and boosts accuracy. Cancer diagnosis and classification using a hybrid Relief F-CNN model was suggested in [16]. The work provides hybrid methods for dimension reduction and classification employing Relief and layered autoencoders. The study used Ovarian, Leukemia, and CNS microarray datasets. Ovarian dataset has 253 samples, 15,154 genes, and 2 classes. Leukemia dataset has 72 samples, 7129 genes, and 2 classes. SVM had the highest classification accuracy without dimension reduction for the ovarian, leukemia, and CNS microarray datasets. Hybrid approach Relief F + CNN method outperformed others. It classified ovarian, leukemia, and CNS datasets with 98.6%, 99.86%, and 83.9% accuracy, respectively. The work recommends that dimensionality reduction can greatly enhanced classification accuracy as seen with the SVM and CNN classifiers. Analysis of Minimum Redundancy Maximum Relevance as a dimensionality reduction strategy for cancer classification using Support Vector Machine classifier was developed [17]. The work used Minimum Redundancy Maximum Relevance (MRMR) as the dimension reduction approach and Support Vector Machine (SVM) as the classifier. PCA was compared with MRMR. Lung cancer and ovarian cancer data tested with MRMR, SVM linear kernel classifier, and polynomial kernel resulted in an F1-score of 1, with 20% of the original feature dataset used for classification. This means the classification was 100% accurate and the system constructed is very good. In colon cancer classification, the F1-score result utilizing MRMR and SVM polynomial kernel classifier was greater than 0.84. It's the same with leukemia cancer classification, where MRMR and SVM polynomial kernel classifiers performed better than classification without dimension reduction (F1-score 0.9657). [18] worked on an efficient attribute selection method for leukemia prediction, and focuses on variable selection techniques by utilizing effective correlation for attribute selection along with ant colony optimization. The work uses ant lion optimization (ALO) in finding optimal feature set which maximizes classification performance. The natural shooting procedure of ant lions is imitated in the proposed ALO algorithm. Support vector machine technique was utilized for the classification of chosen marker genes, giving a comparable performance with existing related system. [19] applied LASSO (least absolute shrinkage and selection operator) to select features. The proposed model was compared with and without LASSO to those of the single CNN and machine learning approaches, such as support vector machines

with radial basis function, linear, and polynomial kernels; artificial neural networks; k-nearest neighbors; bagging trees. The results demonstrate that the suggested model, both with and without LASSO, outperforms existing models. [20] proposed an approach that uses machine learning methods on microarrays of leukemia GSE9476 cells. The primary focus was to foretell the onset of leukemia. The work employed the LDSVM model to classify leukemia in patients and compared its performance with that of the k-fold cross-validation and grid search optimization approaches. Compared to the previous algorithms, the suggested method achieved higher accuracy, precision, recall, and f1 scores. Additionally, a comparative analysis of the outcomes demonstrated the effectiveness of LDSVM. The entire reviewed literatures identified reduced accuracy and low computational speed amidst other shortcomings of the machine learning techniques, which suggest the need for improvement over the milestones in order to archive better performance of the models.

II. Materials and Methods

The method used in this research comprises three subsections which include the feature selection process, model classification, and model performance evaluation. In this study, feature selection was carried out using correlation-based feature selection (CBFS), and classification was done using four algorithms, namely, support vector machine (SVM), K-nearest neighbor (KNN), decision tree (DT), and ensemble method. Lastly, the evaluation of the model was based on accuracy, precision, specificity, sensitivity, and computation time.

A. Feature Selection

Leukemia cancer dataset was acquired from Kaggle repository via figshare.com/articles/dataset/The_microarray_dataset_of_leukemia_cancer_in_csv_format_/13658787. The data contains 3573 attributes and 72 instances of leukemia cancer. Figure 1 shows the dimension of the dataset, viewing the first 5 rows.

In this study, CBFS was implemented in order to select the most salient features for classification. CBFS selects features according to how well they correlate with the target variable. CBFS matches the target sample's class to the class of the centroid closest to the target sample after first calculating the distance between the target sample and each class' centroid. The clearness rating of a feature is determined by the percentage of samples inside that feature that match one another. The second step involves calculating the expected class label using equation 1 for each $x_{i,j}$ in the sample. Following the completion of the calculation of the gap between $x_{i,j}$ and Med (f_j, c_i) for each class, it selects the nearest centroid

$$Med\ (f_{j,s}) \quad (1)$$

where s represents a predicted class label for $x_{i,j}$

The correlation-based feature selection (CBFS) algorithm is a simple filtering technique that uses correlation-based heuristic evaluation functions to arrange subsets in descending order. A suitable subset of attributes, in accordance with the hypothesis, is one that has traits that have a high correlation with the class but that do not correlate with one another. When there is a high correlation between the attribute in question and the other attributes, it is redundant. Features with little association to class should not be taken seriously. It is essential to eliminate characteristics that are redundant and unnecessary. Algorithm 1 represents the CBFS algorithm.

```
data.head()
```

	samples	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V3563	V3564	V3565	V3566	V3567	V3568	V3569	V3570	V3571	Response
0	1	-0.788350	-0.756913	-1.414095	-0.718028	0.473398	3.113805	2.749407	2.628862	3.146849	...	-0.660664	-0.277515	-0.190609	1.096830	0.069212	-0.178846	0.468823	-0.331179	-0.825661	normal
1	2	-1.335163	-1.335163	-1.335163	-1.205542	-0.055226	0.251215	-1.213103	1.040300	3.097184	...	-0.756412	-0.670722	-0.603962	0.263903	0.520380	-0.037259	0.461020	-0.390380	-1.335163	normal
2	3	-1.423499	-1.423499	-1.389461	-0.069438	0.911507	2.080529	1.603549	1.702697	2.980989	...	-0.487601	-0.091597	0.289707	0.328599	0.732303	-0.973264	0.686988	0.355827	-0.708238	normal
3	4	-0.941616	-1.362703	-1.362703	-0.959263	-0.052647	2.210509	1.520901	1.625528	3.244964	...	-1.135454	-0.230745	-0.330132	0.483504	0.590966	-0.852819	0.327239	-0.874228	-1.149951	normal
4	5	-1.373415	-0.527130	-1.373415	-1.191340	0.068572	0.963808	1.654828	-0.319909	3.193077	...	-1.373415	-0.948803	-0.845447	0.306028	0.339066	0.107542	-0.534426	-0.325722	-1.373415	normal

5 rows × 3573 columns

Figure 1: Sample of the Leukemia Dataset

Algorithms 1: Proposed CBFS Algorithm

1 Input F: original feature set
2 N: size of population
3 D: dimension of feature
4 Output S: optimal feature subset
5 Initialize each particle in the population
6 Calculate the matrix R of correlation coefficients between features in F
7 Calculate the contribution of each feature in F by R
8 while The termination condition of the iteration is not satisfied do
9 for i=1 to N do
10 Calculate the fitness value of the particle using KNN classifier
11 Update the historical best position of the particle
12 end for
13 Update the optimal position of the population
14 for i=1 to N do
15 for i=1 to D do
16 Update the velocity of the particle
17 Update the position of the particles combining the w value of each feature
18 end for
19 end for
20 end while
21 Output the optimal position (optimal feature subset)

B. Model classification using SVM, KNN, DT and Ensemble Method

Machine Learning can be an incredibly beneficial tool to uncover hidden insights and predict future trends. In this study, four classifiers were adopted for the implementation of the system.

C. Support Vector Machine (SVM)

SVM can categorize linear and nonlinear data. Each data item is mapped into an n-dimensional feature space. It discovers the hyperplane that splits the data into two classes while maximizing marginal distance and minimizing classification mistakes. The algorithm for SVM implementation is shown in Algorithm 2.

Algorithm 2: SVM [21]

Begin
1. Initialised SVM parameter and structure
2. Generate an initial number of birthing lairs
3. L_l = (f = 1,2,3.....n)
4. While (Stopping criterion)
5. If noise = false
6. Search in the proximity for a new lair by using a Brownian walk
7. Else
8. Expend the search for a way for a new layer by using levy walk
9. End if
10. Evaluate the fitness of each new lair and compare with previous
11. If
12. L^{best,s} > L^{best,k+1}
13. Choose the new lair
14. L^{best} = L^{best,k}
15. Else
16. Go to 4
17. End if
18. Rank the solutions;
19. Return the best lair
20. The global best lair is fed to SVM classifier for training
21. Training the SVM classifier
22. End while
23. End

i. K-Nearest Neighbor (KNN)

kNN is a non-parametric statistical method utilized since the 1970s. kNN retrieves k samples closest to unknown samples in the calibration dataset (e.g., based on distance functions). Calculating the average response variables from k samples determines the label (class) of unknown samples (i.e., the class attributes of the k nearest neighbor). As a result, for this classifier, k is the main tuning parameter for kNN. A bootstrap technique determined k. In this study, k values are tested from 1 to 20 to find the best one for all training samples. Algorithm 3 depicts the implementation steps for KNN algorithm.

Algorithm 3: KNN [22]

```
Nearest-neighbor( D[1..n, 1..n], s )
// Input: A n x n distance matrix D[1..n, 1..n] and an index s of the starting city.
// Output: A list Path of the vertices containing the tour is obtained.
for i ← 1 to n do Visited[i] ← false
Initialize the list Path with s
Visited[s] ← true
Current ← s
for i ← 2 to n do
    Find the lowest element in row current and unmarked column j containing the element.
    Current ← j
    Visited[j] ← true
    Add j to the end of list Path
Add s to the end of list Path
return Path
```

ii. Decision tree (DT)

Decision tree (DT) is a popular machine learning algorithm. It is a tree-like algorithm, where the top node is called the root node, all internal nodes (those with at least one child) reflect input tests. Depending on the test result, the classification algorithm branches to the appropriate child node and repeats until it reaches the leaf node. Decision tree implementation steps are shown in Algorithm 4.

iii. Ensemble method

Ensemble method is a classifier that combines the prediction of two or more base learners for the purpose of generating strong prediction output. Ensemble method algorithm is shown in Algorithm 5.

Algorithm 4: Decision Tree [23]

```
GenDecTree(Sample S, Features F)
Steps:
1. If stopping_condition(S, F) = true then
   a. Leaf = createNode()
   b. leafLabel = classify(s)
   c. return leaf
2. root = createNode()
3. root.test_condition = findBestSpilt(S, F)
4. V = {v | v a possible outcome of root.test_condition}
5. For each value v ∈ V:
   a. S_v = {s | root.test_condition(s) = v and s ∈ S};
   b. Child = TreeGrowth(S_v, F);
   c. Add child as descent of root and label the edge {root → child} as v
6. return root
```

Algorithms 5: Ensemble Classification [24]

Input:

a set of n documents to categorize X={x1, x2,...,xn}
a set of k classifiers C={c1, c2,...,ck}
an user-defined percentage p to form the test set

Declarations:

s is a integer representing the number of documents in the test set (n*p)
class is a matrix of labels: classifiers' labels (k*s)
m is an integer quadratic matrix s*s defined with zeros

Body:

1. Define the test set S using s documents in X
2. Define the training set T with the remaining t documents in X
3. For each ci in C
{
3.1 Train the classifier ci using the categorized documents in T
3.2 Use the trained classifier ci to categorize the documents in S
3.3 Save the resulting labels in class[i,]
}
4. For each o between 1 and s
For each j between 1 and s
For each b between 1 and k
For each i between b+1 and k
IF (class[b,o] == class[i,j])
IF(m[o,j]==0)
m[o,j]=1;
ELSE
m[o,j]=m[o,j]*2;
5. Use m as input of k-means algorithm to form 2 clusters of documents: k1 and k2.
6. Use the SVM-linear algorithm trained on the T set to classify the documents in k1 and k2.
7. The categories corresponding to each cluster are chosen by determining the majority class obtained in each one of them in the previous step.

D. Model Performance Evaluation

In this study, performance of the model is evaluated using the following metrics; accuracy, precision, specificity, sensitivity, precision and recall. Equation to achieve each performance metric is given in equations 2, 3, 4, 5 and 6.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$F-measure = \frac{2*(precision*recall)}{(precision+recall)} \quad (6)$$

Where

TP is true positive, observation is positive, and it is predicted to be positive.
TN is true negative, observation that is correctly predicted to be negative.
FP is false positives, observation that is incorrectly predicted to be positive.
FN is false negatives, observation that is incorrectly predicted to be negative

III. Results and Discussion

A. Results of Feature Selection

The dataset as earlier stated, having 3573 features, it was discovered that the CBFS selected 1220 relevant features from the dataset. The confusion matrix provides insights into the accuracy, precision and recall of the model which can be used to assess its effectiveness. Figure 2 shows the confusion matrix of the CBFS with SVM.

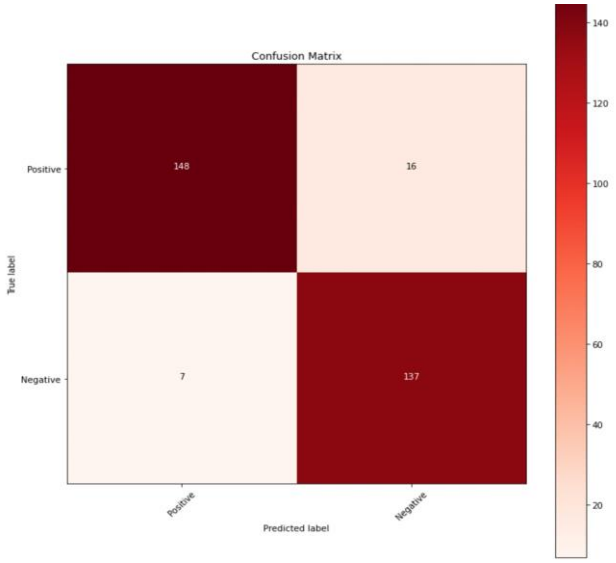


Figure 2: Confusion Matrix of CBFS-SVM (TP=148; TN=137; FP=16; FN=7)

Figure 2 reveals the performance of CBFS-SVM model in predicting leukemia cancer. The model correctly identified 148 positives instances (TP) and 137 negative instances (TN), indicating a good ability to recognize both classes. However, the model incorrectly predicted 16 negative instances as positives (FP), which could lead to further testing. Moreover, it missed 7 positive instances, predicting them as negative (FN), which could equally cause missed prediction. Above all, the confusion matrix suggests that the CBFS-SVM model has good balance of accuracy and error rates. Figure 3 depicts the confusion matrix of SVM only.

Figure 3 reveals the performance of SVM model (without CBFS) in predicting leukemia cancer. The model correctly identified 110 positives instances (TP) and 100 negative instances (TN), indicating a good ability to recognize both classes. However, the model incorrectly predicted 37 negative instances as positives (FP),

which could lead to further testing. Moreover, it missed 17 positive instances, predicting them as negative (FN), which could equally cause missed prediction. Compare to inclusion of CBFS, the confusion matrix suggests that CBFS combination with SVM is significant. Figure 4 depicts the confusion matrix of CBFS with KNN

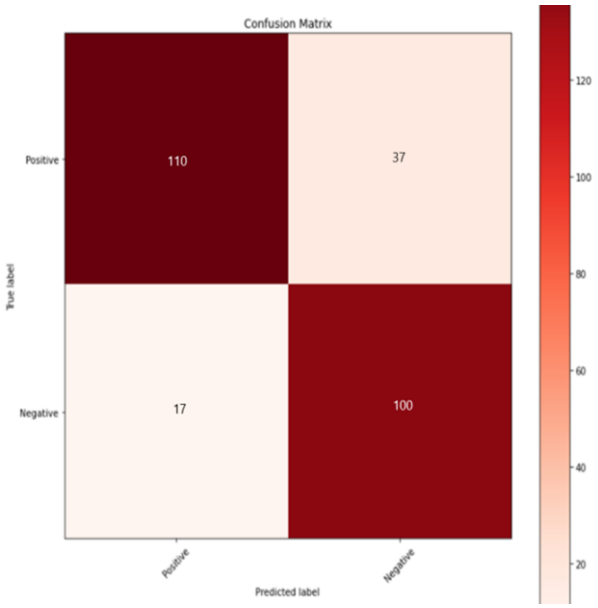


Figure 3: Confusion Matrix of the SVM (TP=110; TN=100; FP=37; FN=17)

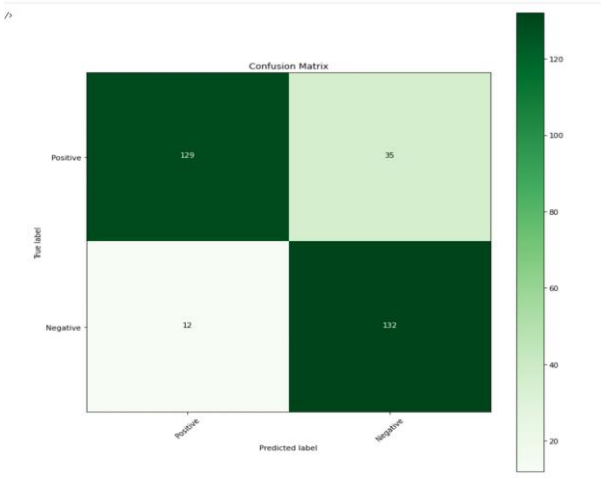


Figure 4: Confusion Matrix of CBFS-KNN Model (TP=132; TN=129; FP=35; FN=12)

Figure 4 reveals the performance of CBFS-KNN model in predicting leukemia cancer. The model correctly identified 132 positives instances (TP) and 129 negative instances (TN), indicating a good ability to recognize both classes. However, the model incorrectly predicted 35 negative instances as positives (FP), which could lead to further testing. Moreover, it missed 12 positive instances, predicting them as negative (FN), which could equally cause missed prediction. Hence, the confusion matrix reinforces that the CBFS-KNN model has good balance of accuracy and error rates.

Figure 5 reveals the performance of KNN model in predicting leukemia cancer. The model correctly identified 107 positives instances (TP) and 100 negative instances (TN), indicating a good ability to recognize both classes. However, the model incorrectly predicted 57 negative instances as positives (FP), which could lead to further testing. Moreover, it missed 44 positive instances, predicting them as negative (FN), which could equally cause missed prediction. This shows that CBFS-KNN model has high error rates.

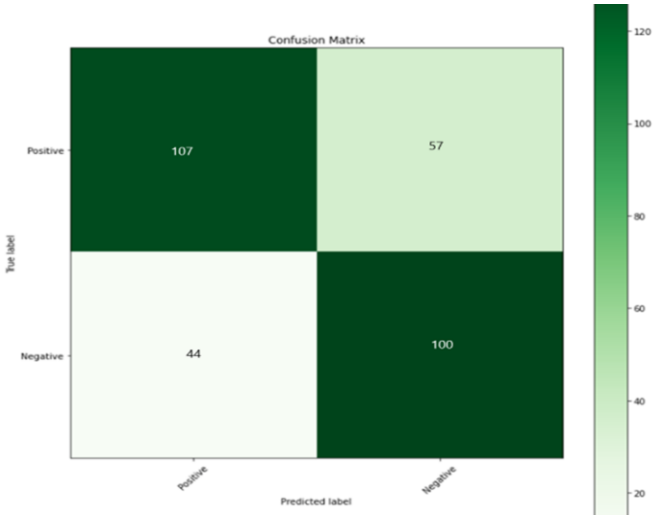


Figure 5: Confusion Matrix of KNN Model (TP=107; TN=100; FP=57; FN=44)

Figure 6 reveals the performance of CBFS-DT model in predicting leukemia cancer. The model correctly identified 160 positives instances (TP) and 138 negative instances (TN), indicating a good ability to recognize both classes. However, the model incorrectly predicted 4 negative instances as positives (FP), which could lead to further testing. Moreover, it missed 6 positive instances, predicting them as negative (FN), which could equally cause missed prediction. Above all, the confusion matrix suggests that the CBFS-DT model has good balance of accuracy and error rates.

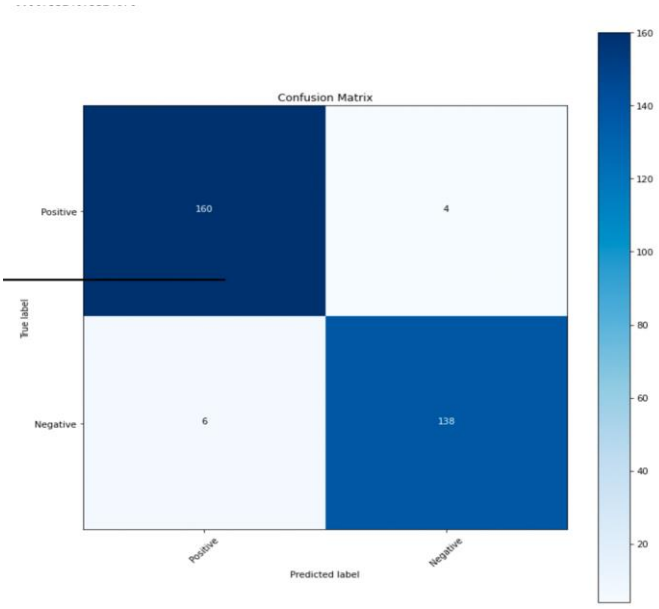


Figure 6: Confusion Matrix of CBFS-DT (TP=160; TN=138; FP=4; FN=6)

Figure 7 reveals the performance of DT model in predicting leukemia cancer. The model correctly identified 149 positives instances (TP) and 118 negative instances (TN), indicating a good ability to recognize both classes. However, the model incorrectly predicted 24 negative instances as positives (FP), which could lead to further testing. Moreover, it missed 36 positive instances, predicting them as negative (FN), which could equally cause missed prediction. The confusion matrix suggests that the DT model has high error rates.

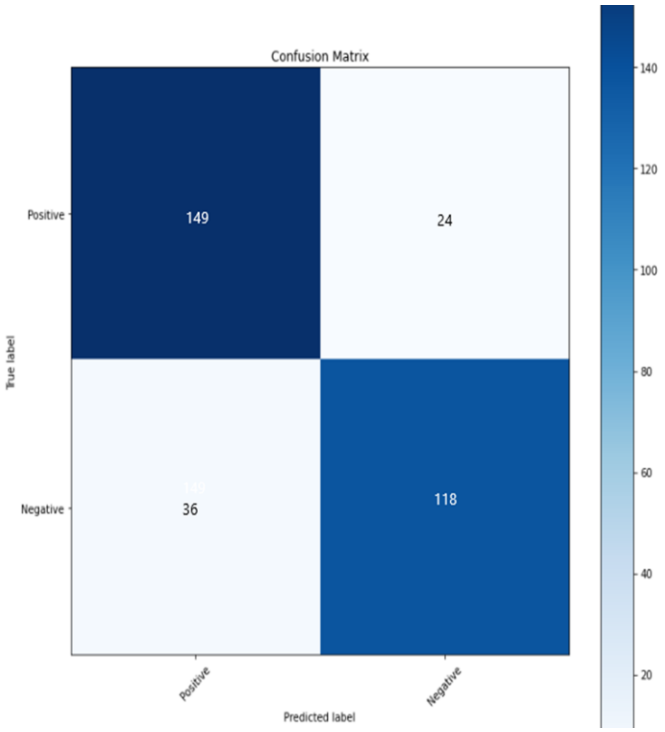


Figure 7: Confusion Matrix of DT (TP=149; TN=118; FP=24; FN=36)

B. Result of Models' Evaluation

In this study, Support Vector Machine, k-Nearest Neighbor, Decision Tree and Ensemble method were employed to classify the models, while CBFS is used to select most relevant features in the leukemia dataset. Performance evaluation measure was obtained using five metrics namely; sensitivity, specificity, precision, accuracy, and F1 score. These metrics were visualized via confusion matrix above. Table 1 reveals that CBFS using DT achieved the highest sensitivity score (96.75%), specificity score (97.18%), precision score (97.56%), accuracy score (96.75%), f1-score (96.97%) and least computational time (0.4336).

Table 1: Performance Comparison of the Classifiers

PERFORMANCE MEASURES	CBFS+ SVM	CBFS+ KNN	CBFS + DT	SVM	KNN	Decision Tree	Ensemble
Sensitivity	95.48	91.49	96.75	86.61	70.86	80.54	67
Specificity	89.54	79.04	97.18	70.87	63.69	83.10	70
Precision	90.24	78.66	97.56	74.83	65.24	86.13	40
Accuracy	92.53	84.74	96.75	78.74	67.21	81.65	70
F1 score	92.79	84.59	96.97	80.29	67.94	833.24	71.33
Computational Time (Sec)	0.5221	1.2632	0.4336	1.4103	1.5332	1.4931	5.9605

The results of the experiment carried out in this research is shown in Table 1. Decision Tree with CBFS outperformed the other models with 96.75% accuracy; however, Support Vector Machine with CBFS followed with accuracy score of 92.53%. Moreover, CBFS+DT model was much faster to execute at less than 1 minute as compared to other models which took approximately 1 minute or more. Hence, it is imperative to note that the performance of the different models can vary significantly based on the data and parameters used.

C. Result of Models' Comparison with Previous Studies

The model accuracy was compared to earlier related work and the result is shown in Table 2. The CBFS+DT model outperformed the existing work based on the recorded accuracy of the model.

Table 4.2: Comparison of the Finding

AUTHOR	TECHNIQUE	RESULTS (Accuracy)
(Dese, Raj et al., 2021) https://doi.org/10.1016/j.clml.2021.06.025	K-Means - SVM	94%
(Mostafa et al., 2022) https://doi.org/10.1186/s12911-022-01980-w	J-48	95%
(Eckardt et al., 2021) https://doi.org/10.1038/s41375-021-01408-w	Deep Learning	92.%
(Almadhor, 2022) https://doi.org/10.3389/fncom.2022.1083649	SVM	90%
(Zhou et al., 2021) https://doi.org/10.3389/fped.2021.693676	CNN	82%
Developed model (2024)	CBFS +DT	96.75%

IV. Conclusion

The results of models' evaluation showed the efficacy of machine learning algorithms in immediate prediction of leukemia cancer. However, the CBFS+DT model has demonstrated a strong performance in predicting leukemia cancer, with a higher number of true positive and true negatives, suggesting that the model is effective in identifying patterns and relationships in the data, and can be relied upon to make accurate predictions and facilitate decision making. In detail, CBFS using DT achieved the highest sensitivity score (96.75%), specificity score (97.18%), precision score (97.56%), accuracy score (96.75%), f1-score (96.97%) and least computational time (0.4336). Nevertheless, the output of the model according to confusion matrix highlights the importance of continued refinement and optimization of the model because the presence of 35 false positives and 12 false negatives indicates that there is still room for improvement. By fine-tuning the feature selection process and adjusting the model's parameters, it is possible to have reduced number of errors and in turn, improve the overall accuracy of the predictions.

References

[1] M. O. Arowolo, O. S. Abdulsalam, I. R. Mope, and G. A. K. Bello, "A Comparative Analysis of Feature Selection and Feature Extraction Models for Classifying Microarray Dataset," Comput. Inf. Syst. J., vol. 22, no. 2, pp. 29-37, 2018.

[2] S. A Monaghan,., J.-L Li,., Y.-C Liu,., M.-Y Ko,., Boyiadzis, M., Chang, T.-Y., Wang, Y.-F., Lee, C.-C., Swerdlow, S. H., & Ko, B.-S. (2022). A Machine Learning Approach to the Classification of Acute Leukemias and Distinction From Nonneoplastic Cytopenias Using Flow Cytometry Data. American Journal of Clinical Pathology, 157(4), 546–553.

[3] R. Kumar, R. Srivastava, & S. Srivastava, (2015). Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features. Journal of Medical Engineering, 2015, 1–14.

[4] Kolawole, M.K., Alaje, A.I., Popoola, A.O. and Bashiru, K.A. (2022).Conceptual Investigation of the Disease Transmission Coefficient in SEIR. Epidemic Model Using Laplace Adomian Decomposition Method (LADM). Vol. 4 No. 1. Pp. 242-249. DOI: 10.36108/ujees/2202.40.0152

[5] M. O. Abolarinwa, A. W. Asaju-Gbolagade, A. A. Adigun, and K. A. Gbolagade, "A Proposed Framework for Face Recognition using Enhanced Local Binary Pattern Algorithm with Chinese Remainder Theorem," UNIOSUN J. Eng. Environ. Sci., vol. 4, no. 2, pp. 54-60, 2022. doi: 10.36108/ujees/2202.40.0260

[6] Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 34(4), 1060-1073.

[7] N. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, and A. Scotto di Freca, "An Experimental Comparison of Feature-Selection and Classification Methods for Microarray Datasets," Information, vol. 10,

no. 3, p. 109, 2019. doi: 10.3390/info10030109

[8] M. S. Al-Batah, B. M. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers," Int. J. Online Biomed. Eng. (IJOE), vol. 15, no. 08, p. 62, 2019. doi: 10.3991/ijoe.v15i08.10617

[9] M. Bukhari, S. Yasmin, S. Sammad, and A. A. Abd El-Latif, "A Deep Learning Framework for Leukemia Cancer Detection in Microscopic Blood Samples Using Squeeze and Excitation Learning," Math. Probl. Eng., vol. 2022, pp. 1-18, 2022. doi: 10.1155/2022/2801227

[10] S, A. A. V., & V, J. P. (2022). A Modified Firefly Deep Ensemble for Microarray Data Classification. *The Computer Journal*. https://doi.org/10.1093/comjnl/bxac143

[11] M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, and A. Roshanpour, "Machine Learning in Detection and Classification of Leukemia Using Smear Blood Images: A Systematic Review," Sci. Program., vol. 2021, pp. 1-14, 2021. doi: 10.1155/

[12] T. Bhadra, S. Mallik, N. Hasan, and Z. Zhao, "Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer," BMC Bioinformatics, vol. 23, no. S3, p. 153, 2022. doi: 10.1186/s12859-022-04678-y

[13] T. Elemam and M. Elshrkawey, "A Highly Discriminative Hybrid Feature Selection Algorithm for Cancer Diagnosis," Sci. World J., vol. 2022, pp. 1-15, 2022. doi: 10.1155/2022/1056490

[14] L.-P. Chen, "Classification and prediction for multi-cancer data with ultrahigh-dimensional gene expressions," PLoS ONE, vol. 17, no. 9, p. e0274440, 2022. doi: 10.1371/journal.pone.0274440

[15] H. Fathi, H. AlSalman, A. Gumaei, I. I. M. Manhrawy, A. G. Hussien, and P. El-Kafrawy, "An Efficient Cancer Classification Model Using Microarray and High-Dimensional Data," Comput. Intell. Neurosci., vol. 2021, pp. 1-14, 2021. doi: 10.1155/2021/7231126

[16] Kilicarslan, S., Adem, K., & Celik, M. (2020). Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network. Medical Hypotheses, 137, 109577.

[17] Ma'ruf, F. A., Adiwijaya, & Wisesty, U. N. (2019). Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier. Journal of Physics: Conference Series, 1192, 012011.

[18] Santhakumar, D., & Logeswari, S. (2020). Efficient attribute selection technique for leukaemia prediction using microarray gene data. *Soft Computing*, *24*(18), 14265–14274.

https://doi.org/10.1007/s00500-020-04793-z

[19] Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., & Omolo, B. (2021). A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Scientific Reports, 11(1), 15626.

[20] Karim, A., Azhari, A., Shahroz, M., Brahim Belhaouri, S., & Mustofa, K. (2022). LDSVM: Leukemia Cancer Classification Using Machine Learning. Computers, Materials & Continua, 71(2), 3887–3903.

[21] Sharif, W., Yanto, I. T. Y., Samsudin, N. A., Deris, M. M., Khan, A., Mushtaq, M. F., & Ashraf, M. (2019). An optimised support vector machine with ringed seal search algorithm for efficient text classification. *Journal of Engineering Science and Technology*, *14*(3), 1601–1613.

[22] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 6256. https://doi.org/10.1038/s41598-022-10358-x

[23] Hambali, M. A., Saheed, Y. K., Oladele, T. O., & Gbolagade, M. D. (2019). Adaboost Ensemble Algorithms for Breast Cancer Classification. Journal of Advances in Computer Research Quarterly, 10(2), 1–10.

[24] Moreira-Matias, L., Mendes-Moreira, J., Gama, J., & Brazdil, P. (2012). *Text Categorization Using an Ensemble Classifier Based on a Mean Co-association Matrix* (pp. 525–539). https://doi.org/10.1007/978-3-642-31537-4_41

[25] Zhou, M., Wu, K., Yu, L., Xu, M., Yang, J., Shen, Q., Liu, B., Shi, L., Wu, S., Dong, B., Wang, H., Yuan, J., Shen, S., & Zhao, L. (2021). Development and Evaluation of a Leukemia Diagnosis System Using Deep Learning in Real Clinical Scenarios. *Frontiers in Pediatrics*, *9*. https://doi.org/10.3389/fped.2021.693676