How to Cite:
Vijayan, S., & Venkatachalam, R. (2022). Classification of acute lymphoblastic leukaemia using machine learning algorithms. *International Journal of Health Sciences*, *6*(S7), 5245-5257.
https://doi.org/10.53730/ijhs.v6nS7.13135

Classification of acute lymphoblastic leukaemia using machine learning algorithms

Saranya Vijayan
Avinashilingam Institute of Home Science and Higher Education for Women,
Dept. of Computer Science, Coimbatore, India
Email: Saranyav101@gmail.com

Radha Venkatachalam
Avinashilingam Institute of Home Science and Higher Education for Women,
Dept. of Computer Science, Coimbatore, India
Email: 19phcsf003@avinuty.ac.in

***Abstract*---**Acute Lymphoblastic Leukemia is a serious disease which may cause death if it is not detected at an early stage. It is common in children as well as adults. The detection of ALL is being done manually by examining the blood smear samples under a microscope. Manual blood testing has got several shortcomings such as it is slow and accuracy is also less. Generally, the inspection is done by an experienced pathologists and if there is any malformations the presence of lymphoblasts will be there. The accuracy of the diagnosis will be dependent on the experience of the operator. The Proposed research work aims in improving the diagnosis of ALL using Machine Learning Classifiers. Few classifiers haven been applied and compared on the segmented dataset images. The automated system can provide several advantages like it will minimize human intervention and it would provide more accurate results. In this research work EDES-SVM and EDSC-SVM have been used for classification. Experimental results obtained are then compared with the results of other machine learning classifiers such as SVM, ESVM, DSC-SVM, DES-SVM. From the experimental results it is analysed that the proposed method outperforms the conventional methods. The results obtained exhibit that the proposed approach could be used to diagnose (ALL) Acute Lymphoblastic Leukemia and its sub-types that will definitely assist pathologists.

***Keywords*---**pathologists, Acute Lymphoblastic Leukemia, EDES-SVM.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 9 May 2022, Manuscript revised: 18 July 2022, Accepted for publication: 27 August 2022

Introduction

Leukemia is caused due to the sudden overproduction of abnormal white blood cells when the white blood cells in the bone marrow increases rapidly. Haematologists have classified leukemia in to different types based on the severity of the disease. The treatment approaches will be decided based on the severity of the disease. In the case of acute Leukemia it will spread more fastly while chronic will be growing slowly and would take longer time to advance. The figure given below shows normal as well as leukemic blood cells.

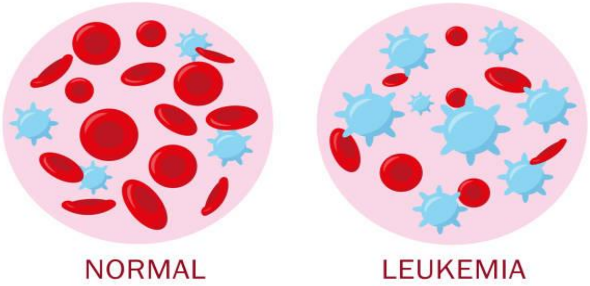


Fig 1. Normal/Leukemic Blood cell

The components of white blood cells are composed of granulocytes and non-granulocytes. Granulocytes comprises of neutrophils, eosinophils and basophils. Non granulocytes has lymphocytes and monocytes in it. White blood cells are a major component of the body’s immune system.

Neutrophils-The main function of Neutrophils is to kill and digest bacteria. They are also responsible for fighting against infections.

Eosinophils- Eosinophils are responsible for killing parasites and cancer cells. It also helps with allergic reactions.

Basophils- When infectious agents enter into the blood ,basophils will give alarm. They secrete chemicals such as histamine.

Lymphocytes- They will create antibodies to fight against bacteria and virus.

Monocytes- The main function of monocytes is to break down bacteria.

The figure below shows the components of a white blood cell in normal and leukemic blood cells.

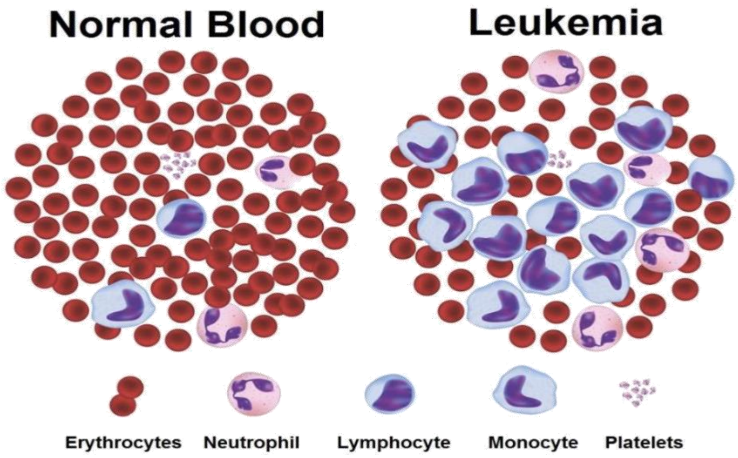
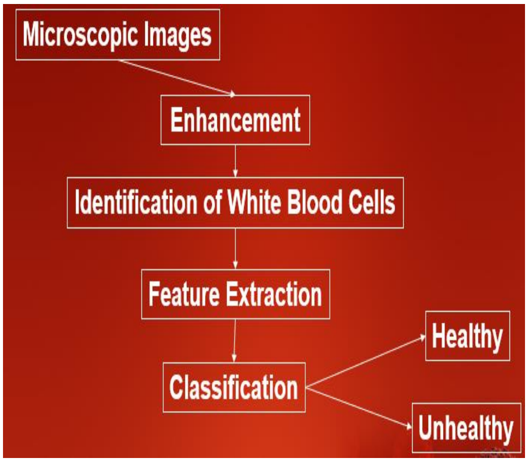


Fig 2. Components of White Blood Cells



Automatic ALL works by acquiring images from the dataset, followed by Image enhancement and Segmentation will be performed to identify the white blood cells. Finally feature extraction and classification will be done to identify the healthy as well as unhealthy blood cells.

In this research work, the Pre processing has been done using some image processing methods such as DWT, K-SVD and UCED. Segmentation has been done using Enhanced K-means and Enhanced Watershed algorithm. The segmented results are then used to classify the identified white blood cells.

Classification is an important part of ALL Identification Classification System, where a number of researches are focusing on improving the classification process. Eventhough, several methods have proved to be successful, recent research results lead to a common agreement that, to a great extent, the creation of a complex classifier that can cover all variabilities inherent to several types of problems and features is problematic. As a solution, the usage of multiple classifier system (or ensemble system) is more viable. The core idea behind multiple classifier methodology is to aggregate a set of learners (known as base classifiers) to obtain a combined classification model, that can produce more accurate results that single learners.

Steps in Ensemble Classification:

There are three stages in ensemble learning process.

-Generation : Generates a set of machine learning models using base classifiers

-Selection : A single best classifier or a subset of best classifiers is (are) selected.

-Integration : Method to combine the results of classifiers from Step 2.

The steps involved are Feature Engineering and Classification.

Feature Engineering consists of two tasks

Feature Extraction - In order to reduce the dimensionality an initial set of the dataset is divided into more groups so that the process will be easier.

Feature Selection- It is the Process of selecting the non redundant and most relevant features.

In the classification step, the feature vector obtained from feature engineering is used to classify a cells as normal or cancerous. If cancerous, to classify them into its types (L1, L2 and L3).

Texture Features (5)

Energy, Entropy, Contrast, Correlation, Homogeneity

Shape Features (10)

Area, Perimeter, Eccentricity, Elongation, Compactness, Minor Axis, Major Axis, Solidity, Form Factor, Nucleus-Cytoplasm Ratio

Color Features (2)

Mean, Standard Deviation

Irregularity of the nucleus boundary (2)

Horizontal Direction

Vertical Direction

Thus, a total of 19 sets of features are extracted

2. Literature Survey

Ruchijogi et al.(2019) have proposed machine learning algorithms for the classification of blood cells. They have created the dataset into 9 classes and the features were extracted to create the feature matrix. They have done a survey on different classifiers and concluded that linear discriminate classifier worked well among all other algorithms.Their proposed classifier had an accuracy of 99% in all the classes.

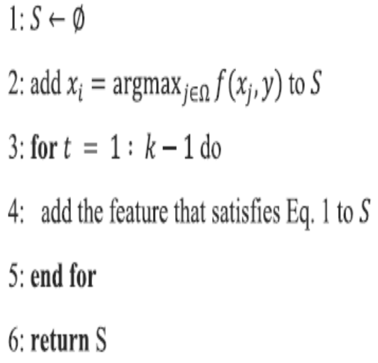
Mehdi Habibzadeh et al. (2018) have investigated a new approach called deep learning approach for the white blood cell detection. They have worked under unfavorable climatic conditions also. In their work they have used different various inception and Resnet deep learning classification. They have mentioned that the future work could be done by using different deep learning algorithms such as convolution deep learning algorithm. Their research work could be extended by making use of their framework in the field of pathological analysis.

A method was developed by Prasidhi G. FalDesai et al. (2018) in which they have detected and classified leukemia in to different types. They have used k means clustering for segmentation. They have acquired an accuracy of 94.7% in classification. The drawback of their algorithm was that they had to pre assign the value of k. Their further research could be extended by focusing on that. The results of their research work can be assessed by using larger dataset

Prayag Tiwari a, Jia Qian b et al. (2019) have proposed convolution deep learning algorithm to detect blood subtypes .The metrics used by them was accuracy. In their research work they did not classify the types of leukemia and that was their limitation.

3.Methodology

This research work proposes the use of a MRMR (Min Redundant Max Relevant) feature selection algorithm to improve classification performance in terms of both accuracy and speed.



```
$S \leftarrow \emptyset$
add $x_i = \text{argmax}_{j \in \Omega} f(x_j, y)$ to $S$
for $t = 1 : k - 1$ do
    add the feature that satisfies Eq. 1 to $S$
end for
return S
```

Ensemble methods have been used in this research work is to get more precise results when we consider multiple scenarios of the same problem. Even though it provides more accurate results there are certain issues associated with it i.e for the creation of an ensemble we need to select a set of classifiers that would provide better predictions. The proposed research work has given a solution to the above mentioned problem by adopting two ensemble models. This research work proposes two ensemble models to tackle the above issue using two methods.

-Dynamic Selection of Classifiers (DSC) - Selects one single best classifier from the set of base classifiers generated

-Dynamic Ensemble Systems (DES) - Selects a subset of best classifiers from the set of base classifiers generated

Proposed solution

-In this research work, DCS and DES are enhanced to improve ALL detection and classification accuracy.

DSC in Ensemble Classification

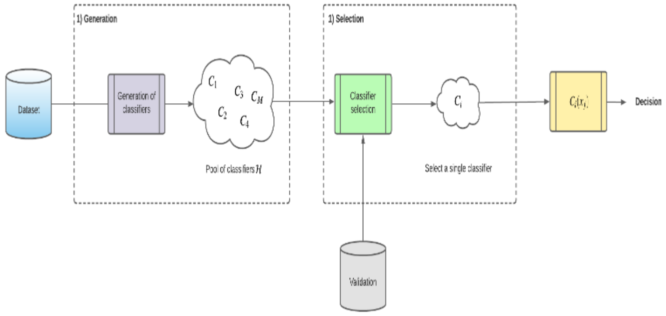


Fig 3.DSC classification

In DSC Ensemble Classification, the best classifier will be selected from a set of classifiers and it will be further used as the classifier.

DES in Ensemble Classification

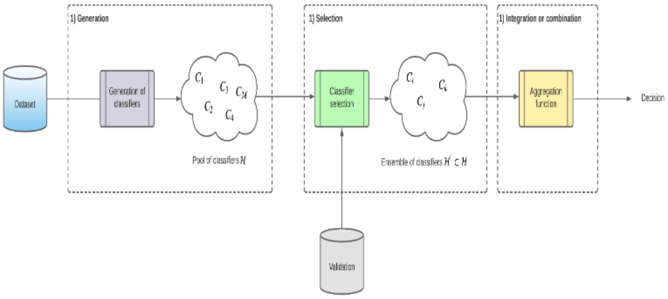


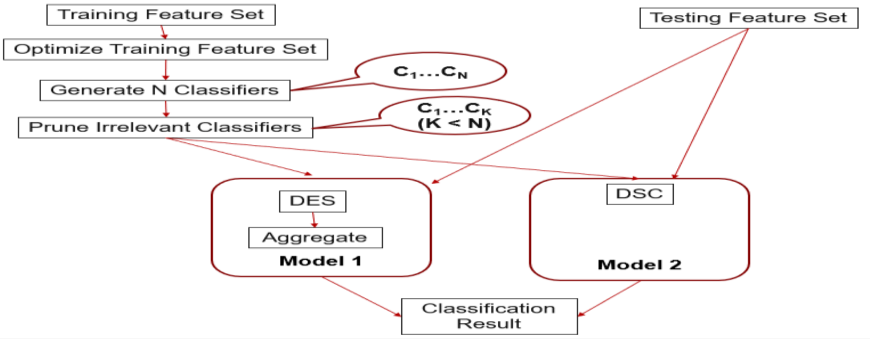
Fig 4. DES classification

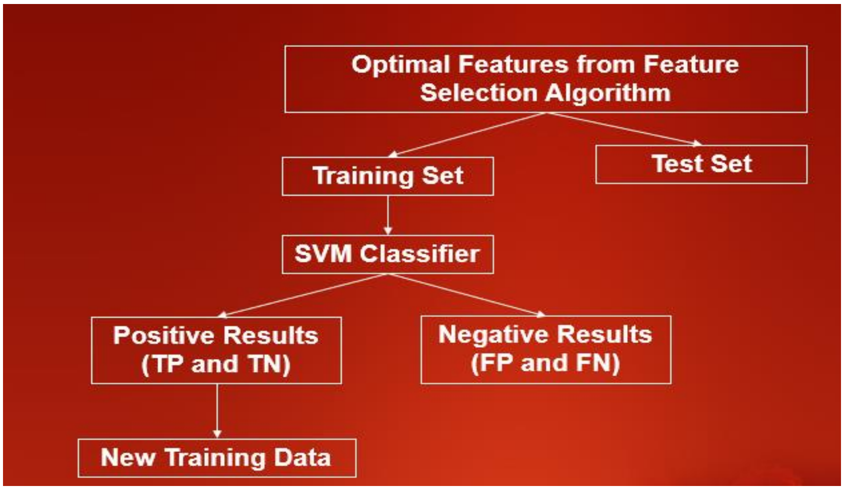
In DES Ensemble Classification, After the generation of a set of classifiers a few classifiers will be selected from that as classifiers and then based on the majority voting the best classifier will be selected from DES as well as DSC for classification.

3.1 Optimization of training set:

Optimization of the training set has been done in such a way that after the features have been extracted, the selection has been done using Minimum Redundant Maximum Relevant (MRMR) algorithm. The dataset has been then divided into Training and Testing set in which the training set is optimized using True positives and True negatives. Bagging method is then used to create 100 different base classifiers.

Pruning algorithm will be applied to eliminate the classifiers that are not needed. The remaining classifiers will be given to DES in which it will select a group of classifiers which work better. Majority voting will be done to select the best classifier and DSC will select only one best classifier.





3.2 Pruning

Pruning algorithm will be applied to eliminate the classifiers that are not needed. The remaining classifiers will be given to DES in which it will select a group of classifiers which work better. Majority voting will be done to select the best classifier and DSC will select only one best classifier.

Need for Pruning Algorithm

-To reduce size of ensemble classification system

-To reduce high memory requirement and computational complexity

The objective of using pruning algorithm is to construct an ensemble system that uses only a smaller subset of base classifiers without compromising on classification accuracy but which can reduce computational complexity.

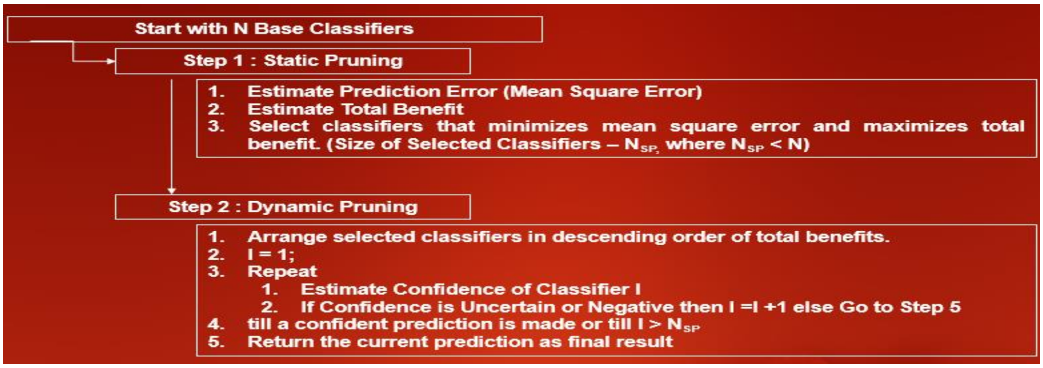
Pruning can be performed using

-Static Methodology : It Selects a subset of base classifiers of fixed size that improves the accuracy with respect to the full ensemble, discarding the rest of them.

Dynamic Methodology: Estimate the number of classifiers required to obtain the final decision for each specific instance during the prediction process. It is much more efficient as it solves both requirements of pruning algorithms

-To further improve pruning process, this research work combines static and dynamic pruning methodologies.

-This proposed method is termed as Pruning Base Classifiers by Combining Static and Dynamic Pruning Algorithm**.**



Start with N Base Classifiers

Step 1 : Static Pruning

1. Estimate Prediction Error (Mean Square Error)
2. Estimate Total Benefit
3. Select classifiers that minimizes mean square error and maximizes total benefit. (Size of Selected Classifiers – $N_{SP,}$ where $N_{SP} < N$)

Step 2 : Dynamic Pruning

1. Arrange selected classifiers in descending order of total benefits.
2. I = 1;
3. Repeat
    1. Estimate Confidence of Classifier I
    2. If Confidence is Uncertain or Negative then I =I +1 else Go to Step 5
4. till a confident prediction is made or till $I > N_{SP}$
5. Return the current prediction as final result

Best Base Classifier Selection Algorithm:

Input

-Set of Base Classifiers A-Classification Accuracy

Output

-Set of best performing classifiers (BHC) that are used to construct heterogeneous ensemble

The Algorithm

```
Step 1 :    BHC = { }
Step 2 :    Arrange base classifiers in S in descending order of its associated accuracy
Step 3 :    if Type = DSC
            BHC - C1
            else
            BHC = BHC + {Cr}
            Return BHC
```

The returned BHC has the best performing classifier set
Details on the design:

Factors	Details
Base Classifier Used	SVM
No. of Base Learning Algorithms (Stage 1)	100
No. of Base Learning Algorithms (Stage 2)	15 (Best Performing Classifier)
Ensemble Creation Methods	Bagging Subspace Selection Technique
Partitioning Method Used	Hold-Out Method
Aggregation Method	Weighted Majority Voting Algorithm

Code	Description
SVM	Single SVM Classifier
ESVM	Ensemble SVM
DSC-SVM	Ensemble SVM using DSC
DES-SVM	Ensemble SVM using DES
EDSC-SVM	Enhanced DSC-SVM
EDES-SVM	Enhanced DES-SVM

4.Results and Discussions

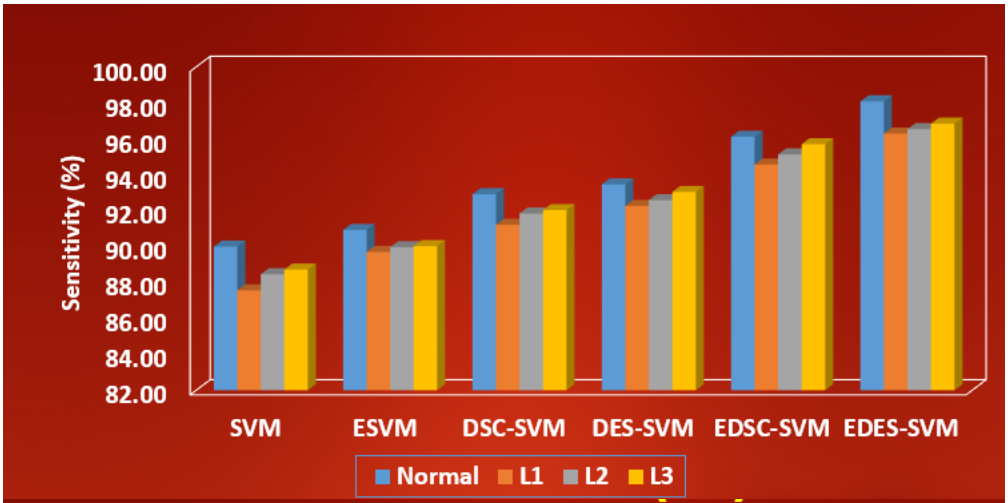
The performance metrics evaluated to assess the performance are,

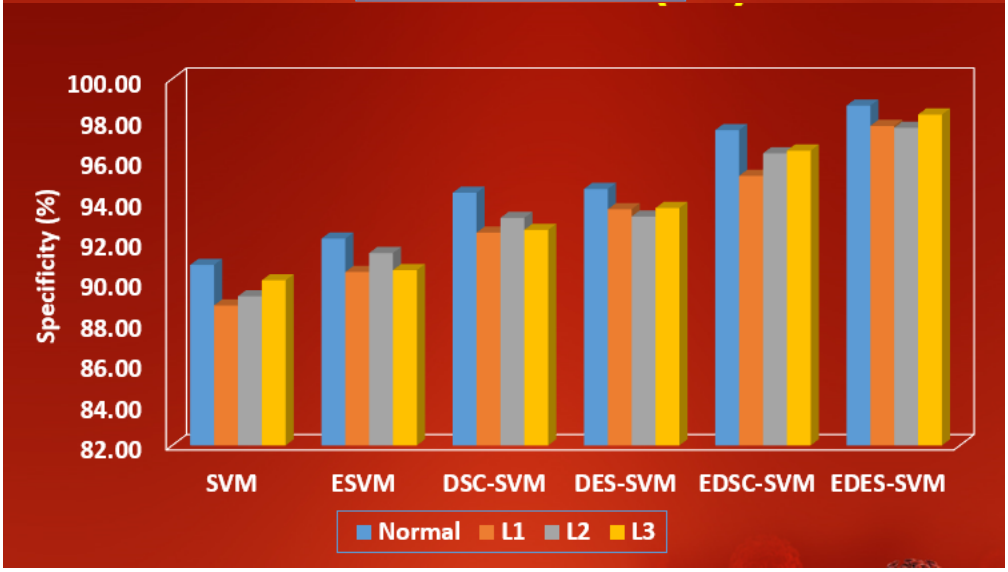
-Sensitivity

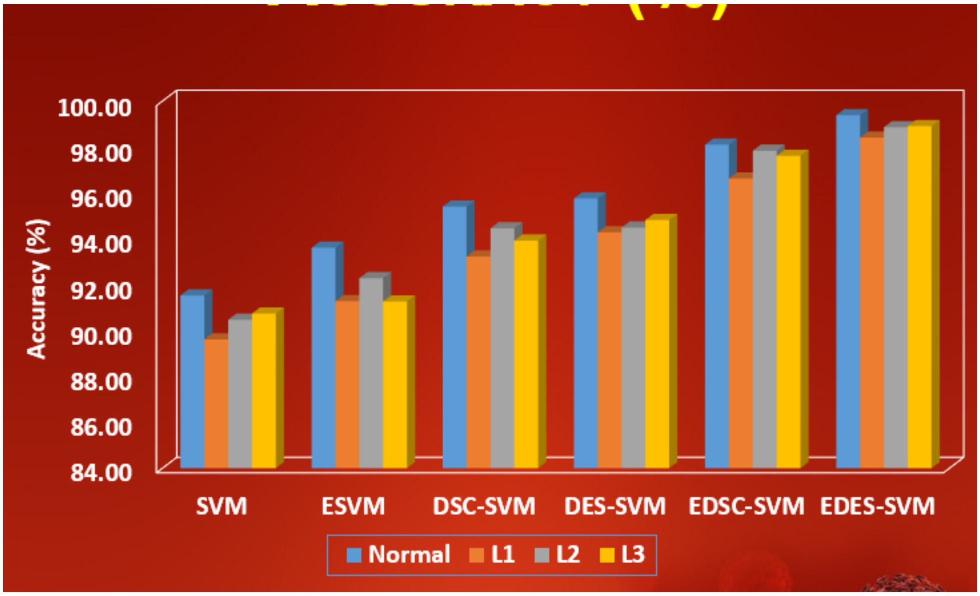
-Specificity

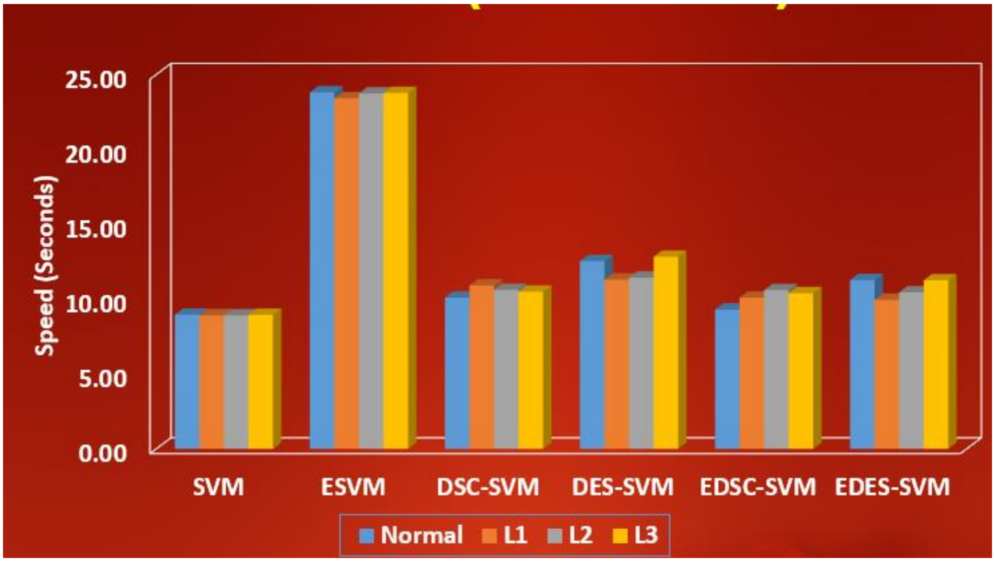
-Accuracy

-Speed





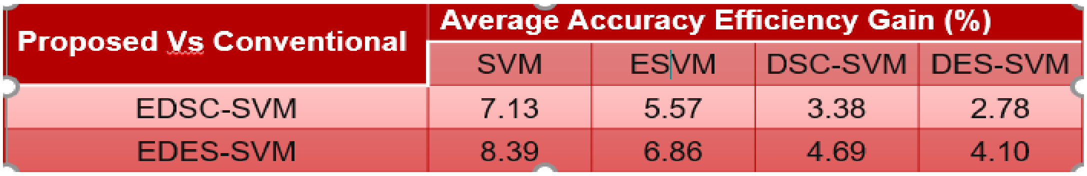




From Phase II experimental results, it can be seen that the proposed ensemble classifiers, EDSC-SVM and EDES-SVM, are more powerful, when compared to single SVM classifier and conventional ensemble systems.

Comparison of the proposed ensemble classifiers showed that the usage of EDES is more beneficial than EDSC.

The efficiency gain of EDES-SVM and EDSC-SVM classifiers over the conventional counterparts are shown in Table.



Proposed Vs Conventional	Average Accuracy Efficiency Gain (%)			
	SVM	ESVM	DSC-SVM	DES-SVM
EDSC-SVM	7.13	5.57	3.38	2.78
EDES-SVM	8.39	6.86	4.69	4.10

The EDES-SVM improved accuracy, on average, by 1.36%, when compared with EDSC-SVM.

Thus, from the results, it can be concluded that the proposed algorithms are efficient in identifying leukaemia efficiently.

Conclusion

This research work has discussed on machine learning classifiers for the classification of ALL and achieved an efficiency gain of 8.39, 6.38, 4.69 and 4.10 when compared with conventional algorithms such as SVM, ESVM, DSC-SVM and DES-SVM with efficient processing time. The research work has proposed two ensemble classifiers namely EDSC-SVM and EDES-SVM and it was found that the proposed algorithms have better accuracy in classification when compared with the traditional counterparts. The proposed approach has acceptable performance, takes the image as input, perform segmentation, and classify as normal if the marrow is not affected or into subtype L1, L2, and L3. The novel contribution of this study is the ensemble classification technique which is not applied before.

References

1. Bhattacharjee, R., & Saini, L. M.(2015). Detection of Acute Lymphoblastic Leukemia using Watershed Transformation Technique. International Conference on Signal Processing, Computing and Control (2015 ISPCC). IEEE Conference,India.
2. Bibin, D.,Nair, M. S., & Punitha, P.(2017). Malaria parasite detection from peripheral bloodsmearimages usingdeep belief networks.IEEEAccess, 5, 9099–9108.
3. Goutam, D., & Sailaja, S. (2015). Classification of acute myelogenous leukemia in blood microscopic images using supervised classifier. International Journal of Engineering Research & Technology (IJERT), 4(1), 569–574.
4. Harouni, M., Rahim, M. S. M., Al-Rodhaan, M., Saba, T., Rehman, A., & Al-Dhelaan, A.(2014). Online Persian/Arabic script classification without contextual information. The Imaging Science Journal, 62(8), 4354.
5. Husham,A.,Alkawaz,M.H., Saba,T.,Rehman,A.,&Alghamdi,J. S.(2016). Automated nuclei segmentation of malignant using level sets.
6. Jagadeesh, S., Nagabhooshanam, E., & Venkatachalam, S. (2013). Image processing based approach to cancer cell prediction in blood samples. International Journal of Technology and Engineering Sciences, 1(1), 1–10.
7. Jamal,A.,Hazim Alkawaz,M.,Rehman,A.,&Saba,T.(2017).Retinal imaging analysis based on vessel detection. Microscopy Research and Technique, 80(7), 799–811.

8. Mughal, B., Muhammad, N., Sharif, M., Rehman, A., & Saba, T. (2018). Removal of pectoral muscle based on topographic map and shapeshifting silhouette. BMC Cancer, 18, 778.
9. Mughal, B., Muhammad, N., Sharif, M., Saba, T., & Rehman, A. (2017). Extraction of breast border and removal of pectoral muscle in wavelet. Biomedical Research, 28(11),5041–5043.
10. Rahim, M. S.M.,Rehman, A., Kurniawan, F., & Saba, T.(2017). Ear biometrics for human classification based on region features mining. Biomedical Research, 28(10), 4660–4664.
11. Rawat, J., Singh, A., Bhadauria, H. S., Virmani, J.,. &., & Devgun, J. S. (2017). Classification of acute lymphoblastic leukaemia using hybrid hierarchical classifiers. Multim