Envisioning the Future of Health Informatics and Digital Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI250058

Extracting Regions of Interest and Selective Feature Application in Leukaemia Image Classification

Marinela BRANESCU[a], Stephen SWIFT[a], Allan TUCKER[a] and Steve COUNSELL[a]
[a]*The Department of Computer Science, Brunel University, West London, United Kingdom*

Abstract: Evaluating the blood smear test images remains the main route of detecting the type of leukaemia, accurate diagnosis is fundamental in providing effective treatment. The changes in the structure of the white blood cells present different morphological characteristics translated into extractable features. This paper explores techniques for manipulating a reduced dataset to increase the classification with CNN (Convolutional neural Network) and feature extraction. Extracting ROI (Regions of Interest) divides the leukaemia images into points of interest respective white blood cells, expanding the dataset an important factor for CNN's performance. Segmenting the initial dataset into ROI through computation after applying Otsu thresholding results in a new dataset of images. The two datasets are analysed, feature extraction performs better on the initial dataset while CNN's accuracy is higher for ROI images. Further steps will divide the images into filtered regions of interest where more specific characteristics are extracted to increase the accuracy.

Keywords: Haralick Texture Features, Feature, Convolutional Neural Network, Otsu Thresholding Method, Regions of Interest

1. Introduction

A cancer of the body's blood-forming tissues and lymphatic system leukaemia affects the white blood cells reducing the body's ability to fight infections [1]. The malignant changes of the white blood cells determine the types of leukaemia being distinguished chronic and acute leukaemia [2]. Chronic leukaemia is characterized by the presence of generally small lymphocytes with scanty cytoplasm and clumped chromatin [6,7]. Acute leukaemia is distinguished by uncontrolled immature blood cells that can have a myeloid or lymphoid lineage [2,6]. In terms of size, the lymphoblast carries smaller cells than the myeloblast, the myeloid type is also defined by the presence of neutrophils in the myeloid line [5,6]. Those morphological and evolutive differences of leukaemia translated in extractable features divide the disease into four main categories Acute Lymphatic Leukaemia (ALL), Chronic Lymphatic Leukaemia (CLL), Acute Myeloid Leukaemia (AML) and Acute Lymphatic Leukaemia (ALL) [1,6]. Developing and affecting the functionality of the blood cells differently detecting the type of disease is fundamental to finding the right treatment [1,2,6].

The white blood cells presenting distinguishable characteristics for each type of

leukemia can be analysed under the microscope to detect the type of disease [2,8]. With the evolution of artificial intelligence and machine learning techniques, CNN have been found to outperform human expertise in medical image classification [1,2,8,]. Feature extraction is an image pre-processing method following several steps for image classification, works by redefining a large set of data into a set of reduced dimensions called features [1,2,6]. Machine learning platforms offer through their classifiers effective tools of analysing the similarities between images by evaluating the extracted features [1,8]. Otsu thresholding technique efficiently separates the blood cells from the plasma in leukaemia images [3]. Combining Otsu with other image augmentation methods will increase the objects visibility and improve the CNN's performance [1,3].

Regions of interest (ROIs) are meaningful and representative information within an image and are considered pointers of the data reliving the most important object for the classification process [6,8]. Extracting ROI can accelerate image processing by avoiding irrelevant image points that are not the targeted object of the classification and can increase a small dataset [7]. Image segmentation is an important area of machine learning image processing, extracting the objects of interest is found to improve the classification accuracy [1,2,4]. In leukaemia image classification white blood cells are fundamental points of interest and dividing the images per white blood cells will increase the dataset and enable the system to train on bigger relevant material [4,6].

The research compares an original dataset of images with a derived expanded dataset resulting from dividing the existing leukaemia images into regions of interest through computation. Both datasets are being analysed with standards CNN packages tested with Cross Validation and traditional feature extraction. The ROI dataset is found to have a higher accuracy for the CNN system due to increasing the dataset and reducing the computation steps by extracting irrelevant pixels from the images. The second experiment where both datasets have texture features, average colours, and size of the cells extracted and analysed with a machine learning platform also is efficient in leukaemia image classification. ROI datasets produce better results at classification for both experiments and dividing the imagers per regions of interest is found efficient.

2. Methods

The initial dataset contains 312 already classified images belonging to the four main types of leukemia ALL, AML, CLL, CM (Figure 1). Otsu is applied on this set of images creating a new dataset where the white blood cells are separated from the plasma increasing their visibility (Figure 2). The new dataset with accentuated points of interest and better visibility of cell's shape prepares the images for extracting regions of interest (Figure 2).

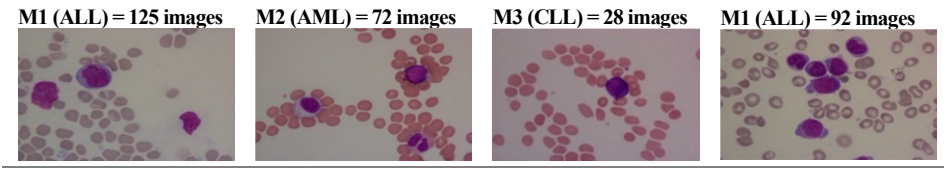

Figure 1: Initial Dataset by leukaemia type showing the number of images for each leukaemia subtype

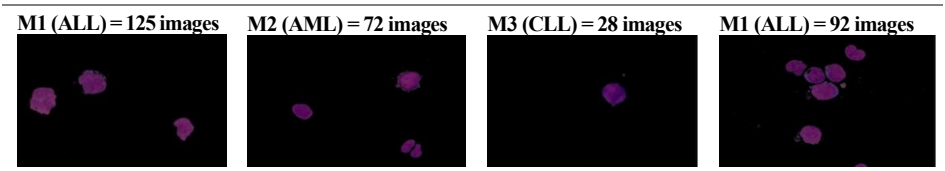

Figure 2: Otsu Dataset resulted after applying Otsu method, by leukaemia type and numbers for each subtype

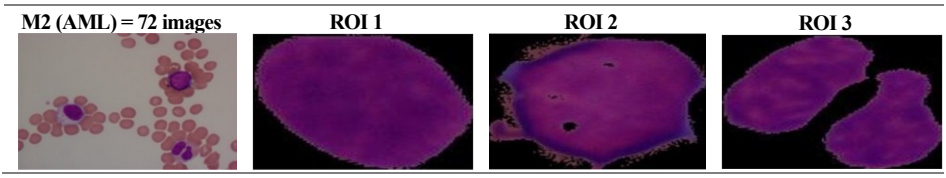

Figure 3: Regions of interest are extracted from the original images creating a new dataset containing one white blood cell per image or a part of it for each leukemia subtype, the above image belongs to AML category

2.1. Experiment I

Training a CNN with standard packages on original and ROI dataset: Both datasets are evaluated with CNN and tested with 5-Fold and 10-Fold Cross-Validation for the instances Drop Out, SMOTE and Shuffle. The computational time is not changed even though the number of images has increased this can be attributed to removing unnecessary information from the image. Is evaluated if the accuracy of the system is improving by dividing the images into regions of interest and if the system is efficient.

2.2 Experiment II

Feature extraction and classification with machine learning platform and validated with computation: The 12 Haralick texture features are extracted through computation from both datasets alongside the average colour, numbers and size of the cells. Selective application of the features on the existing images is tried to determine if random segregation and application of certain features on the images will give different results (Table 1). The 17 features are applied selectively on the ROI dataset creating 131,071 combinations, Yes stands for integrating the feature in the selection and No for omitting the feature. The 12 used Haralick features used represent the Angular Second Moment, Contrast, Correlation, Sum of Squares, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Entropy, Difference Variance, Difference of Entropy, Measure of Correlation.

Table 1. Selective application of feature on the ROI dataset, resulting 131.071 combinations, Feature n is taking values from 2 to 17 and Selection n represents all the possible combinations from Selection 6 to Selection 131,070.

Feature	Selection 1	Selection 2	Selection 3	Selection 4	Selection 5	Selection n	Selection 131,071
Feature 1	Yes	No	No	No	No	Yes	Yes
Feature n	Yes	Yes	Yes	Yes	Yes	No	No

3. Results

As can be seen from looking at the results presented in Table 2, analyzing both datasets with CNN the best results are for ROI dataset for different iterations. The mean accuracy for

the initial dataset noted with Raw (Table 2) being 0.658 while the mean accuracy for ROI dataset being 0.795 making extracting the regions of interest a reliable technique.

Table 2. CNN on Raw Dataset vs ROI Dataset for different iteration instances

Drop Out	Fold CV	SMOTE	Shuffle	Avg.Accuracy Raw	Avg.Accuracy ROI
Yes	10	No	Yes	0.660	0.789
No	10	No	Yes	0.677	0.794
Yes	10	No	No	0.659	0.773
No	10	No	No	0.668	0.780
Yes	5	No	Yes	0.654	0.790
No	5	No	Yes	0.685	0.792
Yes	5	No	No	0.616	0.773
No	5	No	No	0.639	0.781
Yes	10	Yes	Yes	0.660	0.779
No	10	Yes	Yes	0.677	0.774
Yes	10	Yes	No	0.659	0.769
No	10	Yes	No	0.668	0.771
Yes	5	Yes	Yes	0.654	0.775
No	5	Yes	Yes	0.685	0.772
Yes	5	Yes	No	0.616	0.775
No	5	Yes	No	0.639	0.775

Table 3. Extracting features and analyzing it with machine learning platforms for both datasets

Data	Accuracy	ROC Av.	Kappa	Precision	Recall	F-measure
Raw No Cross Validation	0.958	0.998	0.940	0.959	0.958	0.958
Raw 5 Cross Validation	0.907	0.993	0.887	0.907	0.907	0.907
Raw 10 Cross Validation	0.926	0.995	0.894	0.927	0.926	0.926
ROI No Cross Validation	0.881	0.909	0.684	0.880	0.880	0.876
ROI 5 Cross Validation	0.847	0.870	0.595	0.840	0.847	0.840
ROI 10 Cross Validation	0.853	0.873	0.600	0.846	0.853	0.844

Standard feature extraction is more reliable on the Initial Detaset providing better accuracy (Table 3). Selective association of the features applied on the ROI dataset provide higher accuracy for certain combinations, the relevance of features differing (Table 4). The results from analyzing the random selected features with machine learning are evaluated with Computation applied on the extracted features and the difference of accuracy is presented. The insignificant difference of accuracy validates the combination making it trustable when the selected features are extracted and evaluated on both ways.

Table 4. Selective application of the features where 1 means integrating the feature and 0 omitting it

Feature Selection Example	Machine Learning Classification	Computation	Difference
10000011100000000	0.767	0.612	-0.154
10101010010001100	0.701	0.783	0.082
11010000000010000	0.891	0.903	0.012
10000000000000000	0.936	0.944	0.012

4. Discussions

The study presents the importance of increasing the dataset in improving the performance of the CNN reflecting the reliability of deep learning in medical image classification [1,3,8]. The four main types of leukemia starting and developing differently present certain morphological differences that can be expressed in extractable features [1,2,4,5]. The feature's relevance in leukaemia image classification can vary [2,7] and selective application was analysed to determine relevant combinations.

The necessity of increasing the objects visibility before further manipulating the

images was presented in previous studies and Otsu methos was employed to separate the objects from background [3]. Extracting regions of interest is found reliable in similar studies increasing the number of available images for training the CNN and improving the classification's accuracy [1,3,6,7]. The points of interest in leukemia images are white blood cells preserving the targeted characteristics for differentiating the leukemia types [2,5,7]. Comparing the original dataset results with the manipulated set of images provides an insight into the efficiency of adopted techniques and future implementation.

5. Conclusions

The Otsu thresholding method was applied to the initial dataset and is found efficient and accentuating the shapes of the cells [3] and preparing the images for segmentation. The two datasets are evaluated with CNN and feature extraction techniques where in the first instance all the features are applied (Table 3) followed by selective application of the features on ROI Dataset.(Table 4). CNN performs better on the ROI dataset due to increased data and reducing unnecessary pixels, the accuracy of the system being higher than for the initial dataset 0.795 respective 0.658 medium accuracy. Applying the 17 features on both images gives a higher accuracy for the initial dataset but selective features provide a better accuracy for ROI dataset for certain combination of features.

Future work can increase the accuracy of classification by filtering the ROI of extracting images containing irrelevant information like residues of white blood cells and using GAN augmentation.

References

[1] Ahmed IA, Senan EM, Shatnawi HSA, Alkhraisha ZM, Al-Azzam MMA. Hybrid techniques for the diagnosis of acute lymphoblastic leukemia based on fusion of CNN features. Diagnostics. 2023 Feb; 13(6):1026. doi:10.3390/diagnostics13061026.

[2] Haque R, Sakib, AA, Hossain MF, Islam F, Aziz FI, Ahmed FR, Kannan S, Rohan A, Hasan JM. Advancing early leukemia diagnostics: A comprehensive study incorporating image processing and transfer learning. BioMedInformatics. 2024 Jan;4(2):966–991, doi:10.3390/biomedinformatics4020054.

[3] Lam X.-H, Ng K-W, Yoong Y-J, Ng S-B. WBC-based segmentation and classification on microscopic images: A minor improvement. F1000Research. 2021 Nov;10:1168, doi:10.12688/f1000research.73315.1.

[4] Rasheed HH, Abdulazeez AM. Leukemia detection and classification based on machine learning and CNN: A Review. Indonesian Journal of Computer Science. 2024 June;13(3), doi:10.33022/ijcs.v13i3.4044.

[5] Saleem S, Amin J, Sharif M, Mallah GA, Kadry S, Gandomi AH. 2024. Leukemia segmentation and classification: A comprehensive survey. Computers in Biology and Medicine. 2022 Sep;150: 106028, doi:10.1016/j.compbiomed.2022.106028.

[6] Sekar MD, Raj M, Manivannan P. Role of morphology in the diagnosis of acute leukemias: Systematic review. Indian Journal of Medical and Paediatric Oncology. 2023 April;44(05): 464–473, doi:10.1055/s-0043-1764369.

[7] Shahzad M, Ali F, Shirazi SH, Rasheed A, Ahmad A, Shah B, Kwak D. Blood cell image segmentation and classification: A systematic review. PeerJ Computer Science. 2024 Feb;10, doi:10.7717/peerj-cs.1813.

[8] Talaat FM, and Gamel SA. Machine learning in detection and classification of leukemia using C-NMC_LEUKEMIA. Multimedia Tools and Applications. 2023 June;83(3): 8063–8076, doi:10.1007/s11042-023-15923-8.