Comparative Study of Feature Selection for White Blood Cell Differential Counts in Low Resolution Images

Mehdi Habibzadeh, Adam Krzyżak, and Thomas Fevens

Dept. of Computer Science & Software Engineering, Concordia University, Montréal, Québec
{me_habi,krzyzak,fevens}@encs.concordia.ca

Abstract. Features that are widely used in digital image analysis and pattern recognition tasks are from three main categories: shape, intensity, and texture invariant features. For computer-aided diagnosis in medical imaging for many specific types of medical problem, the most effective choice of a subset of these features through feature selection is still an open problem. In this work, we consider the problem of white blood cell (leukocyte) recognition into their five primary types: Neutrophils, Lymphocytes, Eosinophils, Monocytes and Basophils using a Support Vector Machine classifier. For features, we use four main intensity histogram calculations, set of 11 invariant moments, the relative area, co-occurrence and run-length matrices, dual tree complex wavelet transform, Haralick and Tamura features. Global sensitivity analysis using Sobol's RS-HDMR which can deal with independent and dependent input variables is used to assess dominate discriminatory power and the reliability of feature models in presence of high dimensional input feature data to build an efficient feature selection. Both the numerical and empirical results of experiments are compared with forward sequential feature selection. Finally, the results obtained from the preliminary analysis of white blood cell classification are presented in confusion matrices and interpreted using Cohen's kappa (κ) with the classification framework being validated with experiments conducted on poor quality white blood cell images.

1 Introduction and Complete Blood Count (CBC) Interpretation

The examination of peripheral blood smears represents the cornerstone of hematologic diagnosis. Plainly, the examination of the peripheral blood smear is an important indicator of haematological and other abnormal conditions that affect the body of an organism. Blood cells are classified as erythrocytes (Red Blood Cells), leukocytes (White Blood Cells) or platelets (not considered real cells). In all mammals species including human beings, leukocytes, which are less numerous than red blood cells, are divided in two main categories: *granulocytes* and *lymphoid* cells. Granulocyte white blood cell types are Neutrophil, Eosinophil (or acidophil) and Basophil. The lymphoid cells, are separated in Lymphocytes and Monocytes. Expressing the number of white blood cells (WBC) carries many quantitative and informative clues. For example, the increase or decrease of leukocytes is very critical and may prompt detailed medical attention.

The first attempts to build automated laboratory equipment to perform complete blood counts (CBC) was about 60 years ago, in the 1950-1960s [44]. Automatic counting system have been available in the medical laboratories for the last 25 years.

N. El Gayar et al. (Eds.): ANNPR 2014, LNAI 8774, pp. 216–227, 2014.

The instruments used for performing cell counts are based on mix of mechanical, electronic and chemical approaches. Current hematology analyzers used routinely in medical laboratories are such as Sysmex XE-series [35] and also in the Abbott CELL-DYN range [11]. These known systems for white blood cell differential counts reveal good correlation with the manual ground truth reference analysis for neutrophils, lymphocytes, and eosinophils (accuracies of 0.925, 0.922, and 0.877, respectively) and lower accuracy for monocytes and basophils (accuracies of 0.756 and 0.763, respectively). The commonly used approach across biological disciplines and the ground truth is manual WBC counting and type sorting by a trained pathologist or skilled hematology expert, looking at the shape, e.g, nucleus and cytoplasm, occlusion, and degree of contact between cells.

Although the manual inspection method is adequate, it has *three* inevitable types of error: statistical, distributional and also human error [5] such as may happen in poor quality, low magnification view of the slides. Poor magnification and distribution of leukocytes adversely affect the accuracy of the differential count in manual counting. The computerized techniques are the best potential choices to carry out and moderate the load of these regular clinical activities for more efficiency and also to describe the frequency and spatial distribution, and portion of blood smear particles. Hematologists and hematopathologists study and analyze blood smears by looking at cells under an optical microscope. Accordingly, since haematology is a visual science, machine learning and digital image processing have great potential to develop ways to improve haematology research. Computer-aided diagnosis (CAD) also establish methods for accurate, robust and reproducible measurements of blood smear particles status while reducing human error and diminishing the cost of instruments and material used.

In this work, white blood cell analysis of an unfavourable low resolution data set via a feature extraction and selection framework to classify the five mature types of white blood cells is provided. There are no reliable and general comparative studies of feature selection strategies in white blood smear detection with high dimensional input feature data in particular and also in the presence of low quality and unfavourable conditions. This work unifies and extends primary feature vector sets introduced in our earlier work [12, 13], based on using the dual-tree complex wavelet transform (DT-CWT) and few textural features, to high dimensional comprehensive invariant feature sets that also include different invariant shape features such as 11 invariant moments, different histogram calculations, different efficient textual feature such as Tamura and so on. Furthermore, this paper critically examines and compares two feature selection strategies, random sampling-high dimensional model representation (RS-HDMR) and sequential forward selection (SFS), for the white blood cell classification problems in presence of small number of sample set.

2 Background and Literature Survey

The first published paper on blood processing is leukocyte pattern recognition by Bacusmber and Gose in 1972 [2]. In this primary work, classification of white blood cells using some shape features and a multivariate Gaussian classifier into their categories are presented. One decade after, the first fully automated processing of blood smear slides

was introduced by Rowan [34] in 1986. The background on WBC classification by using computer vision concepts is substantial and involves feature extractors, classifiers, quantitative and qualitative process. Ramoser *et al.* [31] used hue, saturation and luminance values to locate WBCs and then leukocytes are classified using a 26-dimensional color feature vector and a classification polynomial support vector machine (SVM). Xiao-min *et al.* [46] introduced method based on threshold segmentation followed by mathematical morphology (TSMM). Sobrevilla *et al.* [40] used fuzzy logic to segment white blood cells from a digital blood smear image. However, in both TSMM [46] and fuzzy logic [40], parameter settings need to set by statistics and experience. Shitong *et al.* [37] proposed white cell detection based on fuzzy cellular neural networks (FCNN). Mukherjee *et al.* [26] proposed a leukocyte detection using image-level sets computed via threshold decomposition. Further, Theera-Umpon *et al.* [43] used four white blood cell nucleus features, and Bayes and artificial neural networks were the classifiers.

Ongun *et al.* [28] proposed an approach using active contours to track the boundaries of white blood cells although occluded cells were not precisely handled. Lezoray [24] introduced region-based white blood cells segmentation using extracted markers (or seeds). Kumar [22] applied a novel cell edge detector while trying to perfectly determine the boundary of the nucleus. Sinha and Ramakrishnan [38] suggested a two-step segmentation framework using k-means clustering of the data mapped to HSV color space and a neural network classifier using shape, color and texture features. Furthermore, in other work, WBC segmentation was achieved by means of mean-shift-based color segmentation in [7] by Comaniciu and Meer while in [19] Jiang *et al.* used watershed segmentation.

Ramesh *et al.* [8] proposed a two-step framework: segmentation and classification of normal white blood cells in peripheral blood smears. Color information and morphological processing were basis functions for segmentation part which was almost close to already published paper in [14]. Latter, WBC classification followed using 19 features such as area, perimeter, convex area, and so on. To lessen the computational burden, Fishers linear discriminant was also applied to trim a multi-dimensional set to six dimensions. In more recent work (2012) Dorini *et al.* [9] introduced automatic differential cell system in two levels to segment WBC nucleus and identify the cytoplasm region. In that work, five mature WBC types were classified using a K-Nearest Neighbor (K-NN) classifier with geometrical shape features and a reasonable accuracy (78% performance vs 85% classified manually by a specialist) was achieved. As a result, despite its long history in cell classification, questions have been raised about the reliability and feature selection in an appropriate white blood cell classification system. On the other hand, one major drawback of these aforementioned approaches is that no general attempt was made to quantify the association between low resolution cell appearance and their classification. Therefore, this latter work would have been more reliable if the framework considered these concerns.

3 Primary Feature Extraction

Continuing previous work [12, 13], the process of feature extraction and parameter estimation is carried out in this extended work. These candidate descriptors have appropriate potential for dealing effectively with challenges and problem in multi-distortion

data set such as blurred, noisy and low magnification of a white blood cell image where internal white blood cell structure is not obvious to detect. All invariant features are scaled to the $[0\ 1]$ range to simplify computational complexity and have consistent inputs for measurement. As a result with all three main feature types in this case, final features vector gives a total of 12140 coefficients for each white blood cell with 28×28 low image size. More details are addressed as below.

Intensity Features: This article examines the mean (μ), standard deviation (σ), skewness (γ_1), and kurtosis (K) in white blood cells classification. However, intensity features may prove inadequate for specially low quality white blood cell data set. A short mathematical background is addressed in our previous research [13]. Eventually, in this case, intensity features gives a total of 788 divided into 784 for raw gray intensity value and 4 measures for histogram calculation feature coefficients for each cell sample.

Shape Features: In terms of pattern recognition, shape descriptors can be classified into two descriptors; *contour-based* and *region-based* shape signifiers. The contour-based descriptors reviewed so far cannot represent ideally white blood cell shapes for which the complete and continuous boundary information is not ideally available with granular and non-uniform borders. Also, questions have been raised about the validity and reliability concern under the constraint of translation, rotation and uniform-scaling invariance properties. In reviewing the literature, the current study found that invariant moment as a region-based calculation which can provide invariant characteristics under different condition are likely occur in translation, changes in scale, also rotation and unique characteristics of a white blood cell that represent its heterogenous shape. Although moment algorithms and theory have been well established in mathematics, far too little attention has been paid to use invariant moment in computer-aided diagnosis (CAD) in medical imaging and for blood smear analysis in particular. This paper has given an account the reasons for the widespread use of *(11)* different invariant moments listed into: M_1 with 332 elements which are moment coefficients for all combined 11 following different moments, $M_2 = 36$ to Radial Tchebichef [27], $M_3 = 36$ to Fourier-Chebyshev magnitude [29], $M_4 = 36$ to Gegenbauer [16], $M_5 = 36$ to Fourier-Mellin magnitude [36], $M_6 = 36$ to Radial Harmonic Fourier magnitude [32], $M_7 = 36$ belong to Generalized Pseudo-Zernike [45], $M_8 = 36$ to Dual Hahn moments [21], $M_9 = 7$ belong to Hu set of invariant moments [17], $M_{10} = 36$ to Krawtchouk [47], $M_{11} = 36$ to Legendre [10, 48], $M_{12} = 1$ to Zernike [25]. In following shape feature category, the relative area (A_r) is also computed [13]. In conclusion, selective shape features provides a total of 333 feature coefficients for each white blood cell sample composed of (332) invariant moment coefficients and one measure for A_r.

Texture Features: This section extends the types of features considered in our earlier work [12, 13]. The vector includes features associated with the Laplace transform, gradient-based, flat texture features [33], and also co-occurrence matrix [15] which is defined over a white blood cell image to be the distribution of co-occurring values at a given offset. Various combinations of the matrix are taken to generate features called *Haralick* features [15] (namely, the angular second moment, contrast, correlation, sum of squares: variance, inverse difference moment, energy, and entropy). Afterwards, six parameters approximating visual perception is used based on the *Tamura* feature [41].

In addition, run-length is an another texture coarseness measurement at typical directions such as 0, 45, 90, and 135 degrees [42]. 11 features for a given gray-level for each individual white blood cell image are extracted. Dual-tree complex wavelet is also examined in this research. It calculates coefficients along rows and columns, and in *six* directions and angles at each individual pixel. The setting, details and proposed framework using DT-CWT is addressed to our previous work [12, 13]. It follows that, textural features gives a total of 11019 feature coefficients for each white blood cell sample. This textural feature vector has been divided into seven aforementioned parts. The first part deals with gradient, Laplacian and flat texture features with 784 for each of them respectively. Then it will go on to Haralick vector and also Tamura textural features with 13 and 6 elements respectively. Finally gray-level run length matrix in four orientations provides 6296 coefficients where dual-tree complex wavelet in six directions also gives a total of 2352 features for each sample.

4 Global Feature Sensitivity and Feature Selection

This work address feature selection algorithm to trace effectiveness of aforementioned high dimensional invariant descriptors in white blood cell classification performance. Feature selection and discriminatory power is achieved using high dimensional model representation (RS-HDMR) and sequential forward feature selection (SFS) along with support vector machine classifier (see section. 5).

RS-HDMR/ Sensitivity Feature Analysis: Lastly, we look at the effect of each individual three multiple features (see section 3) contribution upon the corresponding supervised white blood cell classification. Several studies investigating high-dimensional model representation (HDMR) [1] have been carried out on input and output relationship analysis. High dimensional model representations (HDMR) is a statical approach that depicts the individual or cooperative contributions of the aforementioned features upon the corresponding white blood cell classes. To date, little evidence has been found associating HDMR with image processing and pattern recognition such as Kaya *et al.* research work [20]. Then, future studies on the current topic are therefore recommended. In this work, RS-HDMR approach with a random sample input over the entire domain is used where determination of expansion components is based on shifted Legendre polynomials approximation and Monte Carlo integration [1, 30, 49]. Following that, the influence of individual each input feature variables is computed using global sensitivity approach in which Sobol index is the basis function of calculation [39]. Therefore, global sensitivity indices are denoted by: $S_{i_1,\ldots,i_s}$ where total of the summation $\sum_{s=1}^{n} s_{i_1} + \sum_{1<i<j\leqslant n}^{n} S_{ij}, \ldots + S_{1,2,\ldots,n}$ is equal 1. The first order index S_i is fractional contribution of x_i (each individual feature coefficient) to the variance of $f(x)$ (five main white blood cell classes) where the second order shows the interaction power between x_i and x_j on the classification outcome and these sensitivity analysis terms will be continued. Rabitz *et al.* [1] demonstrated that often the low order interactions of input variables have the dominant impact upon the output assignment. It means that quite often the high ranked global sensitivity feature variable input in mathematical models are first order terms. In the current study, first order S_i for all each individual intensity, shape and texture coefficients are calculated to reach the most effective set.

Sequential Feature Selection: Sequential Feature Selection is an iterative method to select the most informative coefficient by choosing the next feature depending on the already selected features. The method removes redundant and irrelevant features while preserving the efficient features in order to optimize the subset combination of features by considering their predictive efficiency with a given classifier. The method has two distinctive variants: sequential forward selection (SFS), and in contrast, sequential backward selection (SBS) [18] where SFS is taken in this work. In SFS, new added feature x^+ should maximize $J(Y_k + x^+)$ where new component combined with the features Y_k that have already been selected in an iterative and incremental procedure ($x^+ = \arg_{x \notin Y_k} max\ J(Y_k + x^+)$). Despite its simplicity, questions have been raised about the update procedure used by sequential feature selection. For example, SFS is unable to revise optimal feature vector to remove feature variables after the addition of other features. It's also seen that its performance is related to an appropriate criterion to determine the iterative stop point. In this work the optimum criterion value means the minimum error rate in SVM supervised classification where each candidate feature is placed in the new revised subset vector. Several studies investigating SFS have been carried out on medical imaging [4, 6]

5 Discriminant Functions and Support Vector Machine

A linear SVM classifier [3] with 10-fold cross-validation is examined in this work. 10-fold cross-validation is commonly used in presence of a small size (140 samples) of the training and testing data set and with large number of parameters (12140 feature coefficients) to avoid over fitting and to cover all observations for both training and validation. The details of the proposed SVM settings and configuration are addressed in our previous work [12].

6 Experimental Results and Classification

In this section, a set of 140 $8-$bit gray scale poor images with low magnification $(28 * 28)_{px}$ in five balanced dataset (see fig. 1) are used. We have randomly chosen the data to construct the training set after removing almost 20% of the data to be used for testing the SVM classifier.

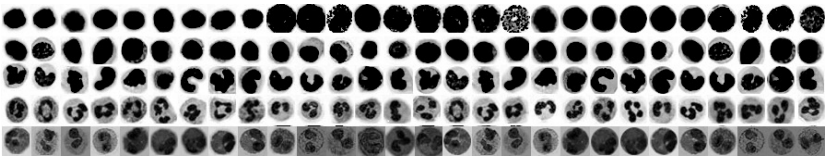

Fig. 1. WBC testing data, each row, top to bottom: Basophil(B), Lymphocyte(L), Monocyte(M), Neutrophil(N), and Eosinophil(E)

Sobol HDMR Analysis: In practice, in the initial configuration for this experiment all 140 samples are used for the RS-HDMR accuracy test. Also, the maximum order for

approximation of the first order $\{f_i(x_i)\}$ terms is 5 where 3 is maximum assigned order for second order $\{f_{ij}(x_i, x_j)\}$. Also a ratio control variate (see section 2.1 in [49]) to supervise and regulate the Monte Carlo integration error with 10 iterations is set for the first and second order RS-HDMR component functions. It also should be noted that in the initial setting to ignore insignificant component functions from the HDMR expansion where the current white blood cell classification system has a high number of input features, a threshold mechanism set to 10% (see section 2.2 in [49]) is also used. During sensitivity analysis, first an intensity feature vector with 788 members composed of 1-784 raw gray scale intensity value, 785 mean, 786 standard deviation, 787 skewness, and 788 kurtosis features is considered. In this case, S_i analysis shows that 38 coefficient out of 788 are computed as non-zero. Sensitivity calculations indicate that indices: 711, 443, 284, 191 and 456 (in range of gray scale intensity value) and 785 (mean value) out of 38 have the first five most discriminative power of $S_i = 0.38$ in this current input - output relationship. Secondly, a shape feature vector with 333 members composed of 1-7 Hu set, 8 Zernike, 9-44 Hahn, 45-80 generalized pseudo-Zernike, 81-116 Chebyshev, 117-152 Legend, 153-188 Krawtchouk, 189-224 Fourier-Mellin, 225-260 Radial Harmonic Fourier, 261-296 Fourier-Chebyshev, 297-332 Gegenbauer, and 333 for relative area is considered. Global Sobol - HDMR Sensitivity calculations demonstrate that 18 of the above feature indices have the highest S_i of 0.82 where in that case, first six indices are: 44 (Hahn coefficient), 191,192 (in range of Fourier-Mellin), 225, 226 (in range of Radial Harmonic Fourier) and 290 (in range of Fourier Chebyshev).

Then a texture feature vector with 11019 members composed of 1-784 gradient, 785-1568 Laplacian, 1569-2352 flat texture, 2352-2365 Haralick texture features, 2365-2371 Tamura, 2372-8667 Gray Level Run Length, and 8667-11019 for dual tree complex wavelet transform features is considered. To provide in-depth analysis of the Sobol index calculation, each of above individual ranges of features is used separately to estimate global sensitivity values. In the case of the gradient features, it can be seen that 43 out of 784 elements have the highest $S_i = 0.44$ where first five indices including 589, 185, 266, 658 and 659 have the most discriminatory power with total $S_i = 0.41$. Next, global sensitivity on the Laplacian features shows that just only 4 elements have non-zero values where these are indices including 421, 309, 337 and 365 with $S_i = 0.17$. Further in flat texture feature analysis result revealed that 13 elements with S_i equal to 0.17 have the dominant power. This suggests that a weak link may exist between Laplacian and flat texture features and the cell classes.

Further, a consequence of the analysis on Haralick features, Tamura shows 9 and 3 with $S_i = 0.7$ and $S_i = 0.6$ have most effective elements in feature - white blood cell class relationship. In terms of Gray Level Run Length feature set, result labeled the subset of 34 elements with $S_i = 0.62$ provides the good predictive power in current HDMR meta-modeling. Global sensitivity in dual tree complex wavelet transform identifies adequate discriminatory power with 111 elements with $S_i = 0.64$ as a major effective subset among all these feature coefficients. In this work based on above explanation 273 elements with exact addressed indices among all 12140 coefficients (almost 2.2%) which are the most convincing set on HDMR input - output relationship in current white blood cell classification system are selected (FV_{HDMR}).

Sequential Feature Selection: For comparison of the results of Sobol HDMR feature selection and to compare the performance on classification accuracy, sequential forward selection (SFS) is used. Sequential forward selection initialized using 10-fold cross-validation by repeatedly calling a criterion based SVM setting (see section 5) with different training and testing subsets of x_{in} and y_{out} where selected feature are saved into a logical matrix in which row (i) indicates the features selected at step (i) with minimum criterion value. In connection with sequential forward selection, many feature indices should be listed here but an exhaustive review is beyond the scope of this current work. Eventually, to do a comparative sensitivity analysis, a feature vector (FV_{SFS}) with the exact number of (FV_{HDMR}) is created. Therefore, this study may leads a difference between classification performance rate (see table 1) for these feature selection algorithms.

Confusion Matrices: A 5×5 confusion matrix is used to represent the different possibilities of the set of instances. The matrices are built on five rows and five columns: Neutrophil; Monocyte; Lymphocyte; Eosinophil; and Basophil representing the known WBC classes whereas for each matrix, each row the values are normalized to sum to 1. Several standard performance terms such as true positive, false positive, true negative, false negative rate, accuracy, precision have been extracted for the confusion matrix. This work addresses kappa (κ) measure as it provides accuracy (AC) versus precision (P) interpretation across class categories [23]. Common Cohen's unweighted κ interpretation is: $\leq 0 \Rightarrow Poor$, $[0, 0.20] \Rightarrow Slight$, $[0.21, 0.40] \Rightarrow Fair$, $[0.41, 0.60] \Rightarrow Moderate$, $[0.61, 0.80] \Rightarrow Substantial$, $[0.81, 1.00] \Rightarrow AlmostPerfect$. The experiments are categorized into set of named selected features (FV_{SFS} and FV_{HDMR}) also with a total high dimensional feature vector with 12140 members (FV_{Total}).

Statistical performance measure is analyzed using analysis of confusion matrices for each named feature & SVM summarized in tables 1a, 1b, and 1c. Further statistical tests revealed that given a small number of input samples (140) in high dimensional feature sets $(= 12140)$ using non-linear SVM kernels leads to over-fitting. The result, as shown in table 1, indicates that for normal low resolution white blood cells using linear SVM & all feature vector FV_{Total} 85% of known white blood cells were classified as such, with this classification rate decreasing to 83% for (FV_{HDMR}) (see table 1c) where the efficiency of (FV_{SFS}) is also 81% which is less than proposed Sobol - HDMR with 83%. RS-HDMR classification performance with 273 elements is less and more similar where classification accuracy is also found with all 12140 coefficients are selected. As confusion matrix tables illustrate, in this poor imaginary database there is not a significant difference between for example the all high dimensional data set and feature selected group with RS-HDMR expansion. The results, as shown in confusion matrix tables indicate that also HDMR results for almost each sub-group is more accurate than SFS method where also sequential forward selection algorithm is too dependent to classifier feedback as well. Also with compare with two ground truth groups, using machines Sysmex XE-series and also Abbott CELL-DYN range (see section 1) it can be seen from the data in confusion matrix tables that global sensitivity with Sobol on RS-HDMR expansion reveals 91% accuracy for Neutrophil, 65% rate for Lymphocyte and also 100% for Eosinophil while the expensive machines mentioned above provide 92.5%, 92.2%, and 87.7%, respectively in an ideal performance. It also provides 81%

classification rate for Monocytes and 77% for Basophils where the results obtained from machines are 75.6% and 76.3%. The following conclusions in regard to κ coefficient can be also drawn from the present confusion matrices. The Cohen's unweighted κ coefficient of the FV_{Total}, FV_{SFS}, also FV_{HDMR} are acceptable (0.81= almost perfect and $0.77, 0.79$ = substantial) in this low resolution WBC classification. Taken together, the most obvious finding to emerge from feature selection and with RS- HDMR study in particular is that all these two methods provide *substantial* performance where lessen computational time and improve model interpret-ability to enhance generalization by reducing over-fitting possibility as well.

Table 1. Confusion matrices (top to down: a,b,c) for SVM classifier, totals over testing images in invariant features & linear SVM

	Linear SVM (FV_{Total}): Assigned WBC classes				
Known	**Basophil**	**Eosinophil**	**Lymphocyte**	**Monocyte**	**Neutrophil**
Basophil	0.72	0	0.21	0.03	0.04
Eosinophil	0	1.00	0	0	0
Lymphocyte	0.17	0	0.68	0.13	0.02
Monocyte	0.01	0	0.04	0.90	0.05
Neutrophil	0	0	0	0.03	0.97

	Linear SVM (FV_{SFS}): Assigned WBC classes				
Known	**Basophil**	**Eosinophil**	**Lymphocyte**	**Monocyte**	**Neutrophil**
Basophil	0.72	0	0.24	0.04	0
Eosinophil	0.00	1.00	0.00	0.00	0.00
Lymphocyte	0.17	0	0.62	0.14	0.07
Monocyte	0.02	0	0.18	0.80	0.0
Neutrophil	0.01	0	0.01	0.04	0.94

	Linear SVM (FV_{HDMR}): Assigned WBC classes				
Known	**Basophil**	**Eosinophil**	**Lymphocyte**	**Monocyte**	**Neutrophil**
Basophil	0.77	0.01	0.17	0.01	0.04
Eosinophil	0	1.00	0	0	0
Lymphocyte	0.16	0.01	0.65	0.1	0.08
Monocyte	0.04	0	0.13	0.81	0.02
Neutrophil	0.02	0.01	0.01	0.05	0.91

7 Conclusions

A machine learning approach for white blood cell classification is effective and reliable, while working under different and even unfavourable and adverse conditions. In this paper, these conditions include low resolution cytological images that are noisy digital white blood cell images. In this research, various approaches to the comprehension and analysis of invariant three main features are presented and the use of these theories is outlined. This work also concentrates on the literature concerning the usefulness of feature selection in presence of big data with high dimensional 12140 invariant features in connection with white blood cell classification. An account is provided of the widespread use of sequential feature selection (SFS) set to recent development in random sample High-dimensional model representation (RS-HDMR). It has conclusively

been shown that these invariant feature collection sets are appropriate solutions as their implementations are promising strategies for representing small distorted white blood cell classifier system (see table 1a). These findings suggest that, in general, RS-HDMR emerged as a reliable input-output relationship predictor of small distorted WBCs and their own classes to allow the full feature sensitivity analysis based on Sobol sequences. It is expected that classification accuracy will be further improved by extending the data set size to reach higher performance in training and testing procedures. The findings are expected to be persuasively supported by future work considering different underdeveloped HDMR variations, i.e., Sobol HDMR using Quasi Monte Carlo, multiple subdomain random sampling HDMR, or Cut-HDMR. Briefly, the empirical findings in this study provide a better understanding of invariant feature implementation and feature selection. One of the more significant findings to emerge from this study is that the possibility of extending the use of this framework to entire field of haematology analysis or other similar medical research.

References

1. Aliş, Ö., Rabitz, H.: Efficient implementation of high dimensional model representations. Journal of Mathematical Chemistry 29(2), 127–142 (2001)
2. Bacusmber, J.W., Gose, E.E.: Leukocyte pattern recognition. IEEE Transactions on Systems, Man and Cybernetics SMC-2(4), 513–526 (1972)
3. Ben-Hur, A., Weston, J.: A user's guide to support vector machines. In: Data Mining Techniques for the Life Sciences, Methods in Molecular Biology
4. Bouatmane, S., Roula, M., Bouridane, A., Al-Maadeed, S.: Round-robin sequential forward selection algorithm for prostate cancer classification and diagnosis using multispectral imagery. Machine Vision and Applications 22(5), 865–878 (2011)
5. Buttarello, M., Plebani, M.: Automated blood cell counts -state of the art. American Journal of Clinical Pathology 130, 104–116 (2008)
6. Choi, K.S., Zeng, Y., Qin, J.: Using sequential floating forward selection algorithm to detect epileptic seizure in EEG signals. In: 11th International Conference on Signal Processing (ICSP), vol. 3, pp. 1637–1640 (2012)
7. Comaniciu, D., Meer, P.: Cell image segmentation for diagnostic pathology. In: Advanced Algorithmic Approaches to Medical Image Segmentation, pp. 541–558. Springer, New York (2002)
8. Dangott, B., Salama, M., Ramesh, N., Tasdizen, T.: Isolation and two-step classification of normal white blood cells in peripheral blood smears. Journal of Pathology Informatics 3(1), 13 (2012)
9. Dorini, L.B., Minetto, R., Leite, N.J.: Semi-automatic white blood cell segmentation based on multiscale analysis. IEEE Journal of Biomedical and Health Informatics 17(1), 250–256 (2013)
10. Fu, B., Zhou, J., Li, Y., Zhang, G., Wang, C.: Image analysis by modified legendre moments. Pattern Recognition 40(2), 691–704 (2007)
11. Grimaldi, E., Scopacasa, F.: Evaluation of the abbott CELL-DYN 4000 hematology analyzer. American Journal of Clinical Pathology 113(4), 497–505 (2000)
12. Habibzadeh, M., Krzyżak, A., Fevens, T.: Analysis of white blood cell differential counts using dual-tree complex wavelet transform and support vector machine classifier. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2012. LNCS, vol. 7594, pp. 414–422. Springer, Heidelberg (2012)

13. Habibzadeh, M., Krzyżak, A., Fevens, T.: Comparative study of shape, intensity and texture features and support vector machine for white blood cell classification. Journal of Theoretical and Applied Computer Science 7, 20–35 (2013)
14. Habibzadeh, M., Krzyżak, A., Fevens, T., Sadr, A.: Counting of RBCs and WBCs in noisy normal blood smear microscopic images. In: SPIE Medical Imaging: Computer-Aided Diagnosis, Orlando, FL, USA, vol. 7963, p. 79633I (February 2011)
15. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3(6), 610–621 (1973)
16. Hosny, K.M.: Image representation using accurate orthogonal gegenbauer moments. Pattern Recognition Letters 32(6), 795–804 (2011)
17. Hu, M.K.: Visual pattern recognition by moment invariants. IEEE Transactions on Information Theory 8(2), 179–187 (1962)
18. Jain, A., Zongker, D.: Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(2), 153–158 (1997)
19. Jiang, K., Liao, Q.M., Dai, S.Y.: A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering. In: IEEE International Conference on Machine Learning and Cybernetics, Xi'an, China, pp. 2820–2825 (November 2003)
20. Kaya, G.T., Kaya, H., Ersoy, O.K.: Feature selection by high dimensional model representation and its application to remote sensing. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4938–4941 (2012)
21. Kok-Swee, S., Faizy Salleh, A., Chee-way, C., Rosli, B., Hock-Ann, G.: Translation and scale invariants of Hahn moments. International Journal of Image and Graphics 09(02), 271–285 (2009)
22. Kumar, B.R., Joseph, D.K., Sreenivas, T.V.: Teager energy based blood cell segmentation. In: 14th International Conference on Digital Signal Processing, Santorini, Greece, pp. 619–622 (July 2002)
23. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33(1)
24. Lezoray, O., Elmoataz, A., Cardot, H., Gougeon, G., Lecluse, M., Elie, H., Revenu, M.: Segmentation of cytological images using color and mathematical morphology. Acta Stereologica 18(1), 1–14 (1999)
25. Li, S., Lee, M.C., Pun, C.M.: Complex Zernike moments features for shape-based image retrieval. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 39(1), 227–237 (2009)
26. Mukherjee, D.P., Ray, N., Acton, S.T.: Level set analysis for leukocyte detection and tracking. IEEE Transactions on Image Processing 13(4), 562–572 (2004)
27. Mukundan, R., Ong, S.H., Lee, P.A.: Image analysis by Tchebichef moments. IEEE Transactions on Image Processing 10(9), 1357–1364 (2001)
28. Ongun, G., Halici, U., Leblebicioglu, K., Atalay, V., Beksac, M., Beksac, S.: Feature extraction and classification of blood cells for an automated differential blood count system. In: International Joint Conference on Neural Networks, Washington, DC, USA, pp. 2461–2466 (July 2001)
29. Ping, Z., Wu, R., Sheng, Y.: Image description with Chebyshev-Fourier moments. Journal of the Optical Society of America A 19(9), 1748–1754 (2002)
30. Rahman, S.: Extended polynomial dimensional decomposition for arbitrary probability distributions. Journal of Engineering Mechanics 135(12), 1439–1451 (2009)
31. Ramoser, H., Laurain, V., Bischof, H., Ecker, R.: Leukocyte segmentation and classification in blood-smear images. In: 27th IEEE Annual Conference Engineering in Medicine and Biology, Shanghai, China, September 1-4, pp. 3371–3374 (2005)

32. Ren, H., Ping, Z., Bo, W., Wu, W., Sheng, Y.: Multidistortion-invariant image recognition with radial harmonic fourier moments. Journal of the Optical Society of America A 20(4), 631–637 (2003)
33. Rodenacker, K., Bengtsson, E.: A feature set for cytometry on digitized microscopic images. Analytical Cellular Pathology 25(1), 1–36 (2001)
34. Rowan, R., England, J.M.: Automated examination of the peripheral blood smear. In: Automation and Quality Assurance in Hematology, ch. 5, pp. 129–177. Blackwell Scientific Oxford (1986)
35. Ruzicka, K., Veitl, M., Thalhammer-Scherrer, R., Schwarzinger, I.: New hematology analyzer Sysmex XE-2100: performance evaluation of a novel white blood cell dierential technology. Archives of Pathology and Laboratory Medicine 125(3), 391–396 (2001)
36. Sheng, Y., Shen, L.: Orthogonal fourier-mellin moments for invariant pattern recognition. Journal of the Optical Society of America A 11(6), 1748–1757 (1994)
37. Shitong, W., Min, W.: A new detection algorithm (NDA) based on fuzzy cellular neural networks for white blood cell detection. IEEE Transactions on Information Technology in Biomedicine 10(1), 5–10 (2006)
38. Sinha, N., Ramakrishnan, A.G.: Automation of differential blood count. In: IEEE International Conference on Convergent Technologies for Asia-Pacific Region, Bangalore, India, pp. 547–551 (October 2003)
39. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation 55(1-3), 271–280 (2001)
40. Sobrevilla, P., Montseny, E., Keller, J.: White blood cell detection in bone marrow images. In: 18th International Conference of the North American Fuzzy Information Processing Societ (NAFIPS), pp. 403–407 (1999)
41. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics 8(6), 460–473 (1978)
42. Tang, X.: Texture information in runlength matrices. IEEE Transactions on Image Processing 7(11), 1602–1609 (1998)
43. Theera-Umpon, N., Dhompongsa, S.: Morphological granulometric features of nucleus in automatic bone marrow white blood cell classification. IEEE Transactions on Information Technology in Biomedicine 11(3), 353–359 (2007)
44. Verso, M.L.: The evolution of blood-counting techniques. Journal of Medical History 8(2), 149–158 (1964)
45. Xia, T., Zhu, H., Shu, H., Haigron, P., Luo, L.: Image description with generalized pseudozernike moments. Journal of the Optical Society of America A 24(1), 50–59 (2007)
46. Xiao-min, Y., Li-min, L., Yu, W.: Automatic classification system for leukocytes in human blood. Journal of Computer Science and Technology 17(2), 130–136 (1994)
47. Yap, P.T., Paramesran, R., Ong, S.H.: Image analysis by Krawtchouk moments. IEEE Transactions on Image Processing 12(11), 1367–1377 (2003)
48. Kang, B., Ma, Z., Ma, J.: Translation and scale invariant of Legendre moments for images retrieval. Journal of Information & Computational Science 8(11), 2221–2229 (2011)
49. Ziehn, T., Tomlin, A.S.: GUI-HDMR - a software tool for global sensitivity analysis of complex models. Environmental Modelling & Software 24(7), 775–785 (2009)