SVM Model for Blood Cell Classification using Interpretable Features Outperforms CNN Based Approaches

William Franz Lamberti
George Mason University / 4400 University Dr
Fairfax, VA 22030
wlamber2@gmu.edu

Abstract

This paper presents a competitive solution for blood cell classification using support vector machines with a polynomial kernel with interpretable features. This approach was able to achieve an overall classification rate of about 98% and outperformed convolutional neural network based approaches by about 5%. Furthermore, by using variables which have a clear meaning, we can create a model which is far more interpretable than its convolution neural network counterparts. The paper and code is available at https://github.com/billyl320/bccd_svm.

Keywords: Shape Classification, Blood Cell, Machine Learning Applications

1 Introduction

A system to classify blood cells are helpful to the health community. Blood cell counts are used to measure the overall health of a given patient and are used to detect a variety of diseases (Cruz et al., 2017). The counting of blood cells is often done by hand, which is a tedious, expensive, and error prone process (Cruz et al., 2017). Thus, an approach which is capable of automating the classification of the various types of blood cells is of great utility.

1.1 Related Work

Convolutional neural networks (CNNs) are a popular modeling technique for classifying images in the computer vision community (Gu et al., 2018). One of the benefits of using a CNN is that they are highly predictive. Unfortunately, this high performance is at the cost of the interpretability of the model (Krizhevsky et al., 2012).

Alam and Islam have presented various ‘you only look once’ (YOLO) CNN models using various architectures (Alam and Islam, 2019). However, none of these models provide a straightforward physical interpretation.

1.2 Contributions

In this paper, we provide a support vector machines (SVM) model with a polynomial kernel using only eight variables to classify blood cell types. This approach provides a more competitive classifier when compared to other modeling approaches. Furthermore, this approach is more interpretable as the features used in the model all have intuitive meanings. The data and code to perform this experiment are provided in the associated GitHub link[1].

2 Methods and Materials

In this section, we will discuss the technical details of our data, metrics used, and SVM model. Since our approach uses interpretable metrics, our model is more understandable than any current CNN based approach.

2.1 Data

We used the publically available Blood Cell Count Dataset (BCCD) (Shenggan, 2019). The provided data had a total of 364 annotated with numerous red blood cells (RBCs), white blood cells (WBCs), and platelets. We first extracted the RBCs, WBCs, and platelets from the provided images using the annotated files. However, a few of the annotations produced errors and were removed. The final counts for the RBCs, WBCs, and platelets were 4153, 372, and 361, respectively.

2.2 Shape Segmentation

We first need to obtain the binary shapes of the blood cells. The entire data set will be passed through a single segmentation algorithm. All of the segmentation algorithms will be described using image operator notation (Kinser, 2018).

The shape segmentation algorithm is

$$\mathbf{b}_i[\vec{x}] = \Gamma_{>125}\mathcal{L}_L\mathbf{a}_i[\vec{x}], \tag{1}$$

where $\mathbf{a}_i[\vec{x}]$ is the input image, $i \in \{1, 2, ..., 4886\}$, $\mathcal{L}_L$ converts the image to grayscale, and Γ is the threshold operator where the intensities greater than 125 are retained.

2.3 Metric Collection

The code to collect the metrics is provided at our GitHub link[1]. The first metrics were the shape proportions, SP, and encircled image-histograms, which is collected from the shape proportion and encircled image-histogram (SPEI) algorithm (Lamberti et al., NA). The

[1] https://github.com/billyl320/bccd_svm

EIs were of particular interest since there is evidence that using them alone generally outperform CNNs in small data scenarios (Lamberti et al., NA). The other shape metrics collected that were used in the model were the eigenvalues of the shapes, eccentricity (Kinser, 2018), circularity (Kinser, 2018), and the number of corners. This results in a total of 8 total metrics used for the SVM model. While additional metrics were collected, they were not needed or were unhelpful for the analysis. They are still provided in the associated GitHub link[1].

2.4 Model

A support vector machines (SVM) with a polynomial kernel model was built. All of the variables were used. The kernel values were determined using 5-folds cross validation on 70% of the data, the training data, using a grid search. The model was then confirmed using the validation data, the remaining 30% of the original data.

3 Results

The confusion matrix for our model is provided in Table 1. A summary of our results and a comparison to other CNN based approaches is provided in Table 2. Our approach has a mean outperformance rate across each classes' corresponding accuracy of 5%.

Table 1: Table shows the confusion matrix for the training and validation data in regular and bolded font, respectively. The columns correspond to the actual class, while the rows correspond to the predicted class. Thus, our model is very accurate on the training and validation data.

Predicted/Truth	Platelet	WBC	RBC
Platelet	249 **(106)**	1 **(1)**	1 **(6)**
WBC	0 **(0)**	93 **(93)**	6 **(3)**
RBC	4 **(2)**	17 **(17)**	2901 **(1236)**

Table 2: Table compares our SVM approach, the first row, to the other CNN approaches attempted by Alam and Islam (Alam and Islam, 2019). Notice that our approach outperforms all of the other approaches for the Platelets and RBC. Our SVM apporach has a mean outperformance rate of about 5%. For the platelet, WBC, and RBC, classes, our approach outperforms by about 18%, -12%, and 10%, respectively.

Approach	Platelet	WBC	RBC
SVM	**98.1%**	83.8%	**99.3%**
Tiny YOLO	96.1%	86.9%	96.4%
VGG-16	73.0%	**100%**	90.9%
ResNet50	79.8%	95.1%	87.3%
InceptionV3	87.8%	**100%**	96.4%
MobileNet	74.2%	93.4%	83.6%

4 Discussion

The SVM model was able to provide competitive results to the CNN approach. With a mean outperformance rate of about 5%, our approach provides competitive results. While it does underperform for classifying WBC observations, it is able to greatly outperform for the RBC and Platelet classes. We believe that this is due in part to the inclusion of EIs. Using the EIs in conjunction with other interpretable metrics allowed for a model that described the complex shapes of blood cells to be created.

5 Conclusions

A SVM model provides competitive results based on interpretable variables compared to CNN based approaches. We believe that this is in part due to the inclusion of EIs which are known to on average outperform CNNs by themselves. Thus, by using EIs in conjunction with other interpretable metrics, we are able to build models that explain more complex shapes.

6 Funding

The Office of the Provost at George Mason University funded this research.

References

Mohammad Mahmudul Alam and Mohammad Tariqul Islam. 2019. Machine learning approach of automatic identification and counting of blood cells. *Healthcare Technology Letters*, 6(4):103–108, July.

Dela Cruz, C. Jennifer, Valiente, Leonardo C. Castor, Celine Margaret T. Mendoza, B. Arvin Jay, L. Song Cherry Jane, and P. Torres Bailey Brian. 2017. Determination of blood components (WBCs, RBCs, and Platelets) count in microscopic images using image processing and analysis. In *2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, pages 1–7, December. ISSN: null.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2018. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, May.

Jason M. Kinser. 2018. *Image Operators: Image Processing in Python*. CRC Press, Boca Raton, FL, 1st edition, October.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

William F. Lamberti, Jason Kinser, and Michael Eagle. N/A. Shape Proportions and Encircled Image-Histograms Improve Analysis and Classification of Shapes for Small Data. *Under Review*.

Shenggan. 2019. Shenggan/BCCD_dataset, November. original-date: 2017-12-07T11:54:25Z.