Electronic Supplementary Material (ESI) for Lab on a Chip.

Supplementary Materials for

Assessing Red Blood Cell Deformability using Deep Learning

Erik S. Lamoureux, Emel Islamzada, Matthew V. J. Wiens,

Kerryn Matthews, Simon P. Duffy, Hongshen Ma

Correspondence to: Hongshen Ma hongma@mech.ubc.ca

For consideration in *Lab on a Chip*

This PDF file includes:

Fig. S1. Microfluidic device manufacturing validation

Fig. S1 displays the microfluidic device sorting results from 14 separate device trials using 1.53µm polystyrene beads. Eight tests were conducted by User 1 and six tests by User 2. Overall sorting outcomes between the users were consistent, possessing similar means ($\mu_{user\ 1}$ = 2.60 and $\mu_{user\ 2}$ = 2.56) and standard deviations ($\sigma_{user\ 1}$ = 0.25 and $\sigma_{user\ 2}$ = 0.19). Using a two-sample student t-test assuming unequal variances with two tails we get p = 0.72. This result indicates that there is no significant difference in sorting results between the two users. In addition, the range of data is constrained, indicating intra-user sorting consistency, in addition to inter-user consistency. User 1 has conducted previous validated experiments with this device[1] while User 2 conducted the microfluidic experiments in this study.

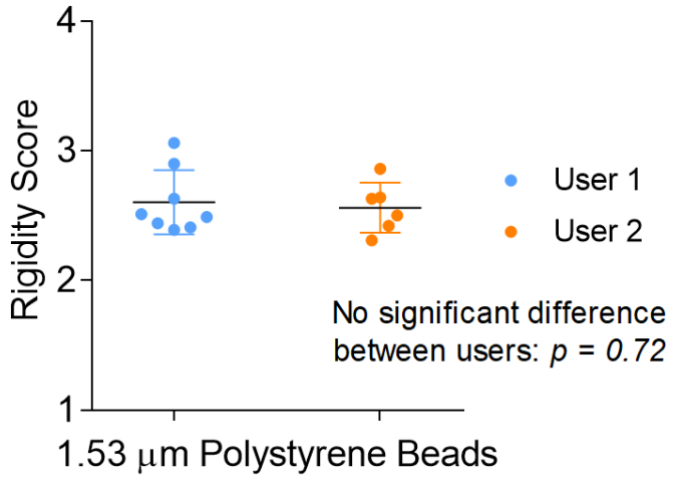


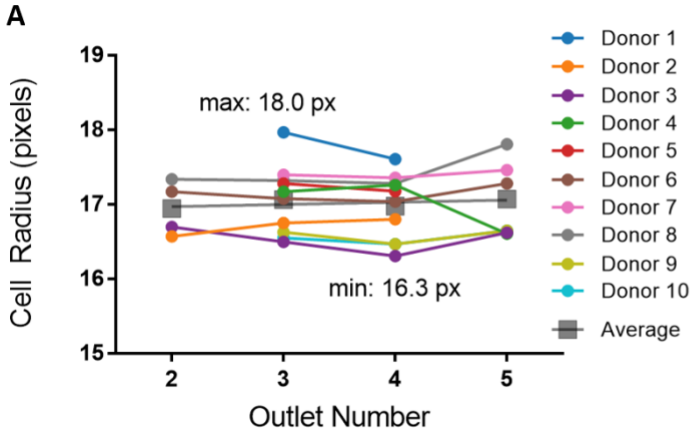
Fig. S1. Microfluidic device manufacturing validation.

Fig. S2. Hough circle transform analysis

To assess whether cell diameter or size is related to cell deformability sorting or to deep learning outcomes, a Hough circle transform analysis was conducted in Python 3.7 using the OpenCV package. The RBC radii for donors by deformability outlet is plotted in **Fig. S3A**, with the averages ranging from 16.3 to 18.0 pixels. From the figure, we observe that RBC cell radius is unique to each donor. This is supported by an analysis of variance (ANOVA),

finding that there were significant differences ($p < 0.0001$) in cell radii between donors (**Fig. S3B**). Therefore, the size of RBCs varies between donors in our dataset, although with radii differences averaging in the single-digit pixel range.

In addition, **Fig. S3A** shows that, in most cases, a donor's cell radii are roughly constant across the outlets. This is demonstrated by an ANOVA applied on the cell radii for each outlet. We find that there was no significant difference in average cell radii for the different deformability outlets ($p = 0.87$) (**Fig. S3B**). As such, we can conclude that the imaged planar cell radius does not substantially contribute to deep learning classification. Therefore, our model learns and detects cell morphological features other than cell size.



B

Analysis of Variance (ANOVA)		
Comparison	Cell radius differences between outlets	Cell radius differences between donors
Significance	None, $p = 0.87$	High, $p < 0.0001$

Fig. S2. Hough circle transform-determined cell radii for all donors separated by outlets (**A**) and analysis of variance (ANOVA) tests (**B**).

Fig. S3. Accuracy degradation by downsampling

We conducted deep learning using training and testing image data that was downsampled to simulate the use of data acquired at lower microscope magnifications. Downsampling 40× images is an imperfect representation of imaging at lower magnifications but is a fine approximation. We find that minor accuracy degradation occurs at low levels of downsampling, but the overall trends of the accuracies converge to random chance (50%) as

the downsampling factor is increased (**Fig. S3**). This relationship is as we expect, indicating there are no major artefacts influencing learning.

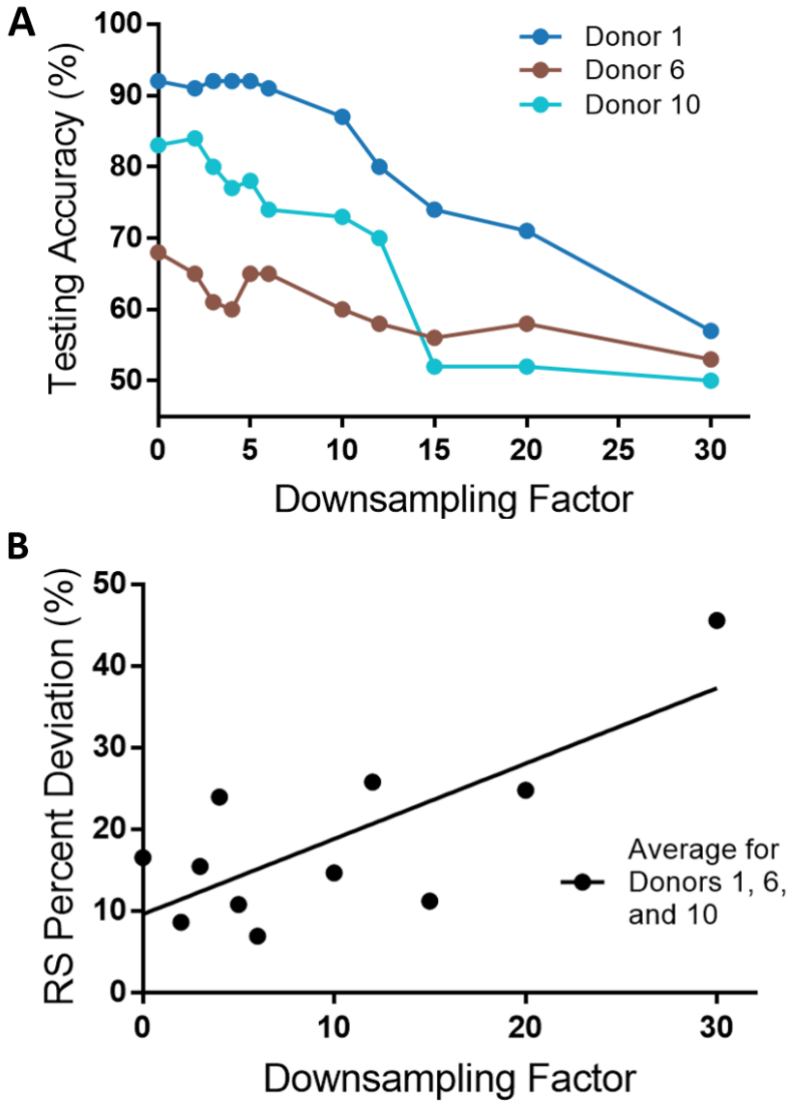


Fig. S3. Testing accuracy (**A**) and RS percent deviation (**B**) degradation by downsampling images.

Fig. S4. Accuracy degradation by reducing training set size

We trained the deep learning model on donors 2, 7, and 10 in seven trials using 10,000, 5,000, 2,000, 1,000, 500, 200, and 100 cell images per class. The resulting models were tested on a test set 20% of the size of the training set. The model's accuracy degrades when

training with fewer cells per class. We find similar trends between the three selected donors, with at least 5,000 images needed per class for reliable testing outcomes.

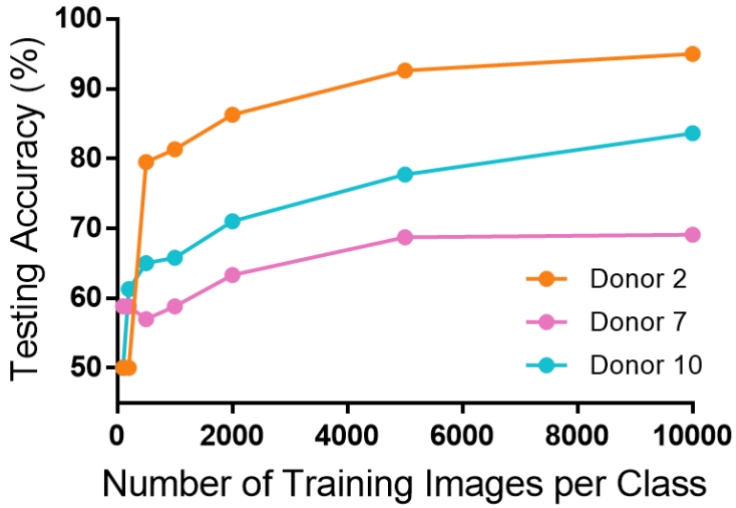


Fig. S4. Sensitivity analysis. Testing accuracy plotted against average testing accuracy for select donors (2, 7, and 10).

Table S1. Additional deep learning results

Table S1 provides additional deep learning results including testing accuracy, precision, recall, F1-score, and the area under the curve (AUC) of the receiver operating characteristics (ROC) curve. These metrics are defined as

$$accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$precision = \frac{TP}{(TP + FP)}$$

$$recall = \frac{TP}{(TP + FN)}$$

$$F1 - score = 2\frac{(precision \times recall)}{(precision + recall)}$$

where

$$TP \equiv True\ positive\ proportion$$

$$FP \equiv False\ positive\ proportion$$

$$TN \equiv True\ negative\ proportion$$

$$FN \equiv False\ negative\ proportion.$$

In summary, precision is the proportion of positive identifications that were correct while recall is the proportion of actual positives that were identified correctly. It is often the case that precision and recall are oppositely related, such that improving one metric may reduce the other. Therefore, the F1-score metric can be used to represent a harmonic mean between precision and recall scores. Finally, the receiver operating characteristic (ROC) curve is a plot showing the performance of the model at different classification thresholds. This plot consists of the True Positive Rate (recall) on the y-axis and the False Positive Rate on the x-axis. An overall assessment of the ROC curve can be obtained using the area under the curve (AUC) metric that, as the name suggests, measures the area under this ROC curve. AUC values range from 0 to 1, with 0 indicating all predictions are incorrect and 1 indicating all predictions are correct. These additional metrics, along with testing accuracy, are displayed in **Table S1**. To a large extent, the precision, recall, F1-scores, and AUC metrics all reflect the testing accuracy of the model.

Table S1. Additional deep learning results including accuracy, precision, recall, F1-score, and AUC of ROC curve.

Donor	Testing Accuracy	Precision	Recall	F1-Score	Receiver Operating Characteristic (ROC) Curve
		(Class 1 – *deformable*, Class 2 – *rigid*)			Area under the curve (AUC)
1	0.92	(0.92, 0.92)	(0.92, 0.92)	(0.92, 0.92)	0.98
2	0.95	(0.96, 0.95)	(0.95, 0.96)	(0.95, 0.95)	0.99
3	0.84	(0.78, 0.91)	(0.92, 0.75)	(0.85, 0.82)	0.94
4	0.83	(1.00, 0.74)	(0.65, 1.00)	(0.79, 0.85)	0.94
5	0.82	(0.87, 0.79)	(0.76, 0.88)	(0.81, 0.83)	0.91
6	0.64	(0.65, 0.63)	(0.61, 0.67)	(0.63, 0.65)	0.69
7	0.71	(0.72, 0.70)	(0.69, 0.74)	(0.71, 0.72)	0.79
8	0.67	(0.63, 0.72)	(0.79, 0.54)	(0.70, 0.62)	0.76
9	0.95	(0.93, 0.97)	(0.97, 0.93)	(0.95, 0.95)	0.99
10	0.81	(0.80, 0.82)	(0.82, 0.79)	(0.81, 0.80)	0.90

References

1 E. Islamzada, *Lab on a Chip*, 2020, 11.