Detection of rare circulating tumor cell clusters in blood using correlation-based feature engineering of light scattering data and machine learning

Vishwesh Srinivasan, Georgios Georgalis, Pramesh Singh
Data Intensive Studies Center, Tufts University

ABSTRACT

Metastasis is a critical stage in tumor progression, involving the spread of tumor cells to form secondary tumors. In this study, we utilized in vivo flow cytometry (IVFC) data and machine learning techniques to model the relationship between light scattering and fluorescence signals for detecting circulating tumor cell (CTC) clusters. By analyzing data from detected CTC clusters, we identified wavelength combinations showing significant correlation differences between cluster and non-cluster conditions. Using correlation-based feature engineering, we constructed a labeled dataset and trained a random forest classifier. The optimized model achieved strong performance. However, applying the model to an entire dataset from a single experiment revealed challenges, especially with false positive predictions leading to lower precision.

INTRODUCTION

Metastasis is the process by which the tumor cells from the primary tumor spread to distant sites, where they form secondary tumors.

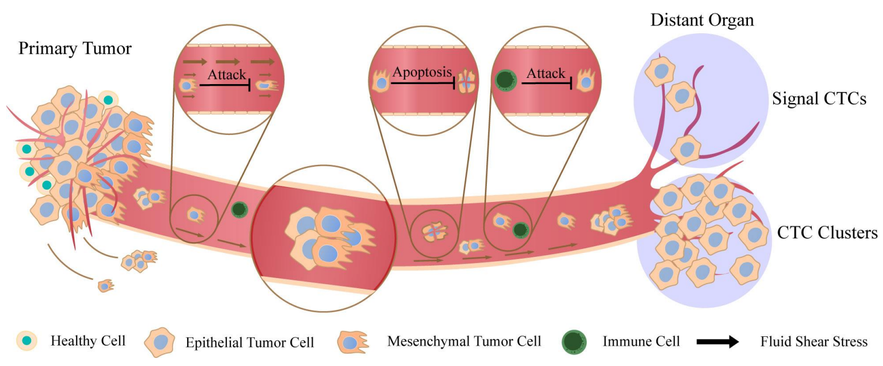


Figure 1: Metastasis process depicting single CTCs and CTC clusters (Chen et al., 2022)

Circulating tumor cell (CTC) clusters are 23–50 times more metastatic than single CTCs (Aceto et al., 2014). In vivo flow cytometry (IVFC) is used to detect circulating tumor cell clusters within live animals or patients. This approach involves labeling CTC clusters with fluorescent markers specific to tumor cells, allowing for their identification. Fluorescently labeled cells for IVFC have limitations in clinical translatability in human subjects.

In this study, we explore the use of machine learning techniques to model the relationship between light scattering and fluorescence data. The data was previously collected through flow cytometry studies, in which the blood samples spiked with GFP-expressing CTC clusters were exposed to three wavelengths: 405 nm, 488 nm, and 633 nm (Vora et al., 2022).

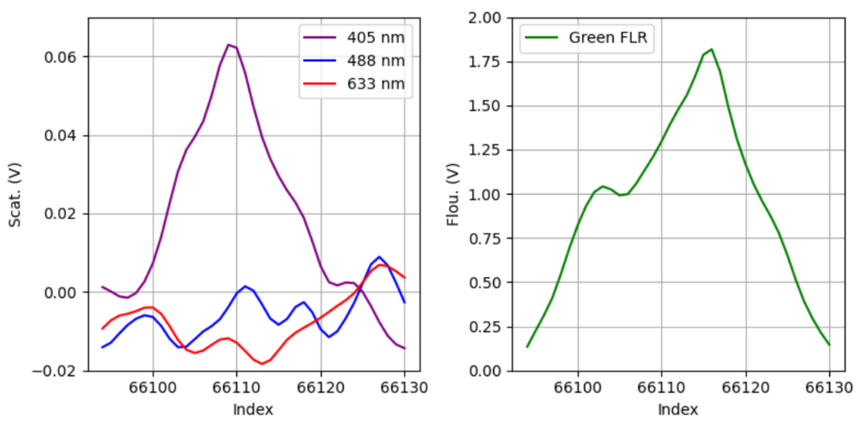


Figure 2: Light scattering and fluorescence data for a 37 data points CTC cluster

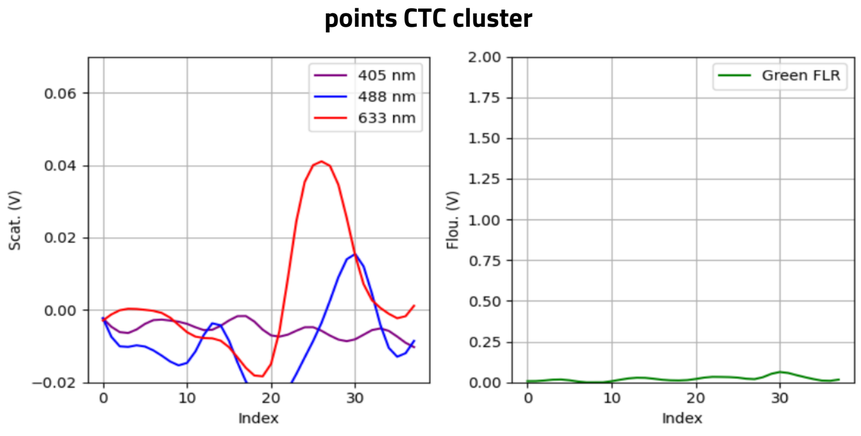


Figure 3: Light scattering and fluorescence data for 37 data points without CTC clusters

FEATURE ENGINEERING AND MODELING

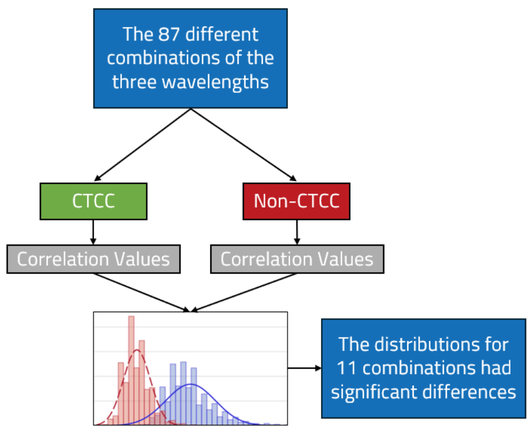


Figure 4: Flowchart depicting the correlation analysis performed.

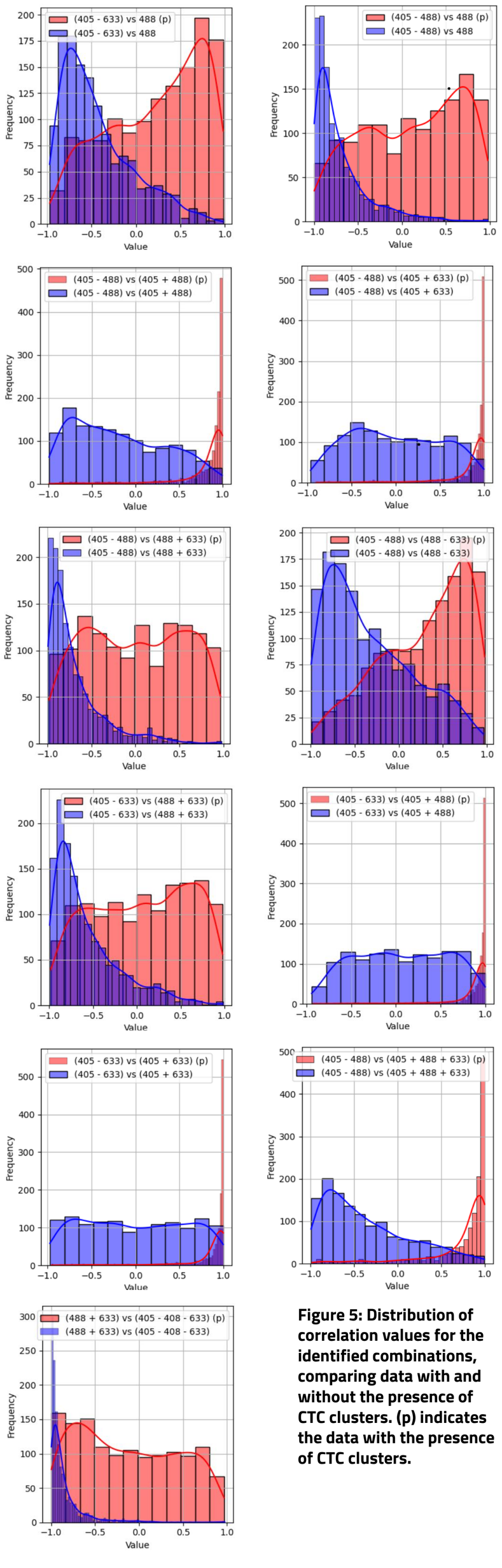


Figure 5: Distribution of correlation values for the identified combinations, comparing data with and without the presence of CTC clusters. (p) indicates the data with the presence of CTC clusters.

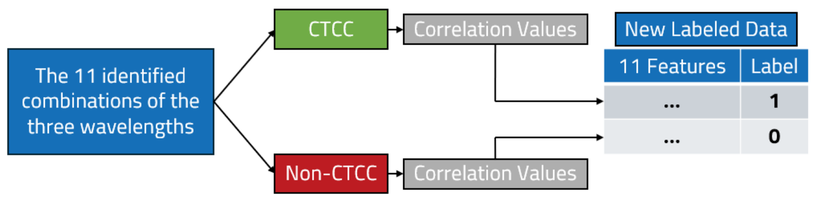


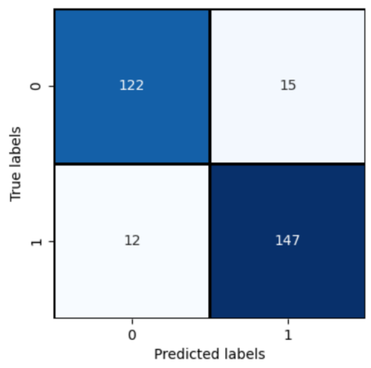
Figure 6: Flowchart depicting the correlation-based feature engineering.

Using the new labeled dataset, we trained a random forest classifier with hyperparameter tuning via a grid search method combined with stratified k-fold cross-validation (5 splits), optimizing for the area under the ROC curve (AUC) as the scoring metric. The parameter grid included a range of 'n_estimators' (from 100 to 3000) and 'max_depth' (from 2 to 20) values to identify the most optimal model configuration.

RESULTS

The optimal model, evaluated on separate testing data not used during training, achieved the following performance metrics.

AUC	Accuracy	TPR	TNR	Precision
0.91	90.88%	92.45%	89.05%	90.74%

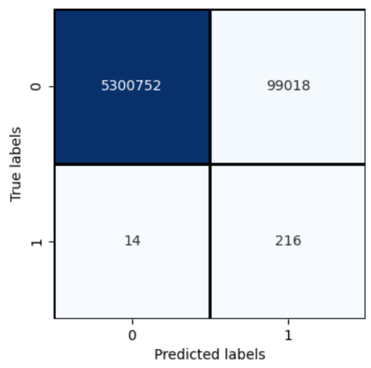


The high AUC and balanced performance across sensitivity, specificity, accuracy, and precision underscore the model's robustness and potential utility.

Figure 6: Confusion matrix obtained for the testing data.

To further evaluate the model's performance, we applied it to the entirety of data obtained from a single experiment comprising approximately 5.4 million data points. Within this dataset, an average of 100 to 200 data points corresponded to the presence of CTC clusters. We grouped the data into segments of 18 data points, calculated correlation values for the 11 combinations of wavelengths, and got the predictions using the model. The following are the metrics obtained.

AUC	Accuracy	TPR	TNR	Precision
0.96	98.17%	93.91%	98.17%	0.22%



All metrics, except for precision, exhibit consistent ranges. However, the model shows a notable number of false positives, leading to a precision (purity) ranging from 0.02% to 0.22%.

Figure 7: Confusion matrix obtained for an entire single experiment data.

REFERENCES

Aceto, N., Bardia, A., Miyamoto, D. T., Donaldson, M. C., Wittner, B. S., Spencer, J. A., Yu, M., Pely, A., Engstrom, A., Zhu, H., Brannigan, B. W., Kapur, R., Stott, S. L., Shioda, T., Ramaswamy, S., Ting, D. T., Lin, C. P., Toner, M., Haber, D. A., & Maheswaran, S. (2014). Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell, 158(5), 1110–1122. https://doi.org/10.1016/j.cell.2014.07.013

Chen, Q., Zou, J., He, Y., Pan, Y., Yang, G., Zhao, H., ... & Lu, Y. (2022). A narrative review of circulating tumor cells clusters: A key morphology of cancer cells in circulation promote hematogenous metastasis. Frontiers in Oncology, 12, 944487.

Vora, N., Shekhar, P., Esmail, M., Patra, A., & Georgakoudi, I. (2022). Label-free flow cytometry of rare circulating tumor cell clusters in whole blood. Scientific Reports, 12(1), 10721.

ACKNOWLEDGMENTS

We extend our sincere gratitude to Irene Georgakoudi and her laboratory team for providing us with the flow cytometry data utilized in this study.

Tufts University | Data Intensive Studies Center