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. We use these correlation-based features to train a random forest classifier for identifying CTC clusters. The optimized model performed well in detecting the presence and absence of CTC clusters from the scattering data, highlighting its potential for clinical applications in tumor diagnostics and monitoring.

INTRODUCTION

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

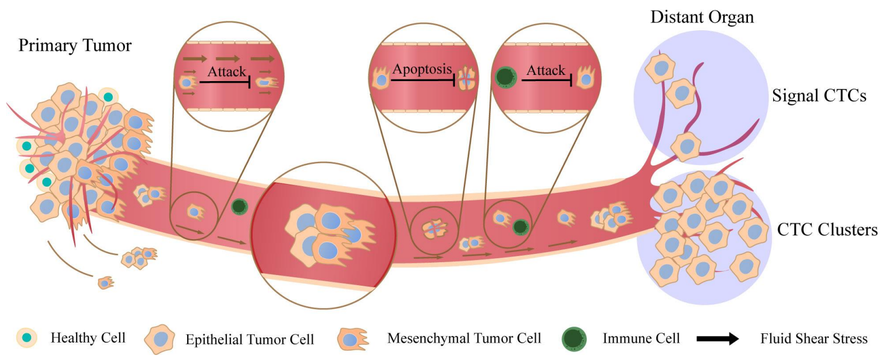


Figure 1: Metastasis process depicting single CTCs and CTC clusters (Chen et al., 2022)

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 excitation wavelengths: 405 nm, 488 nm, and 633 nm (Vora et al., 2022).

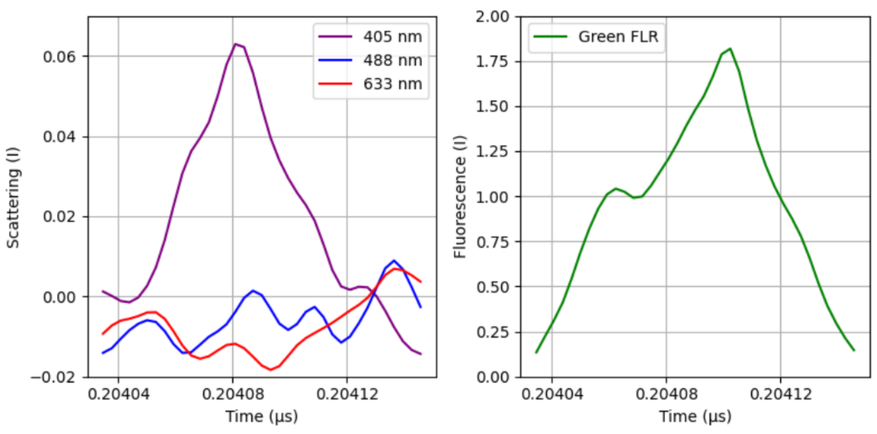


Figure 2: Light scattering and fluorescence data for a 37 data points CTC cluster

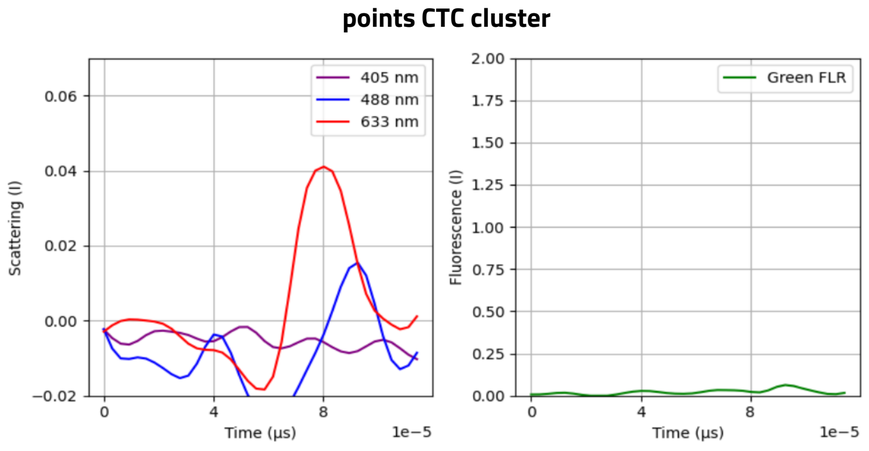


Figure 3: Light scattering and fluorescence data for 37 data points without CTC clusters

FEATURE ENGINEERING AND MODELING

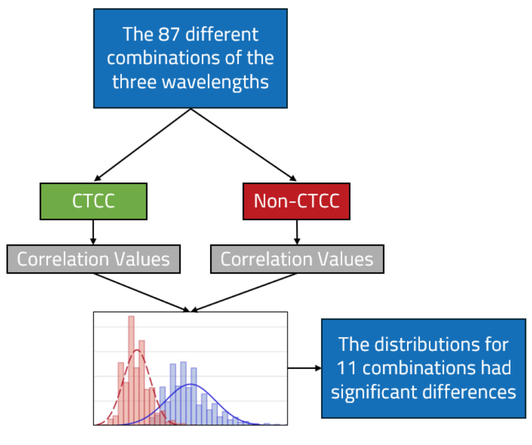


Figure 4: Flowchart depicting the correlation analysis performed.

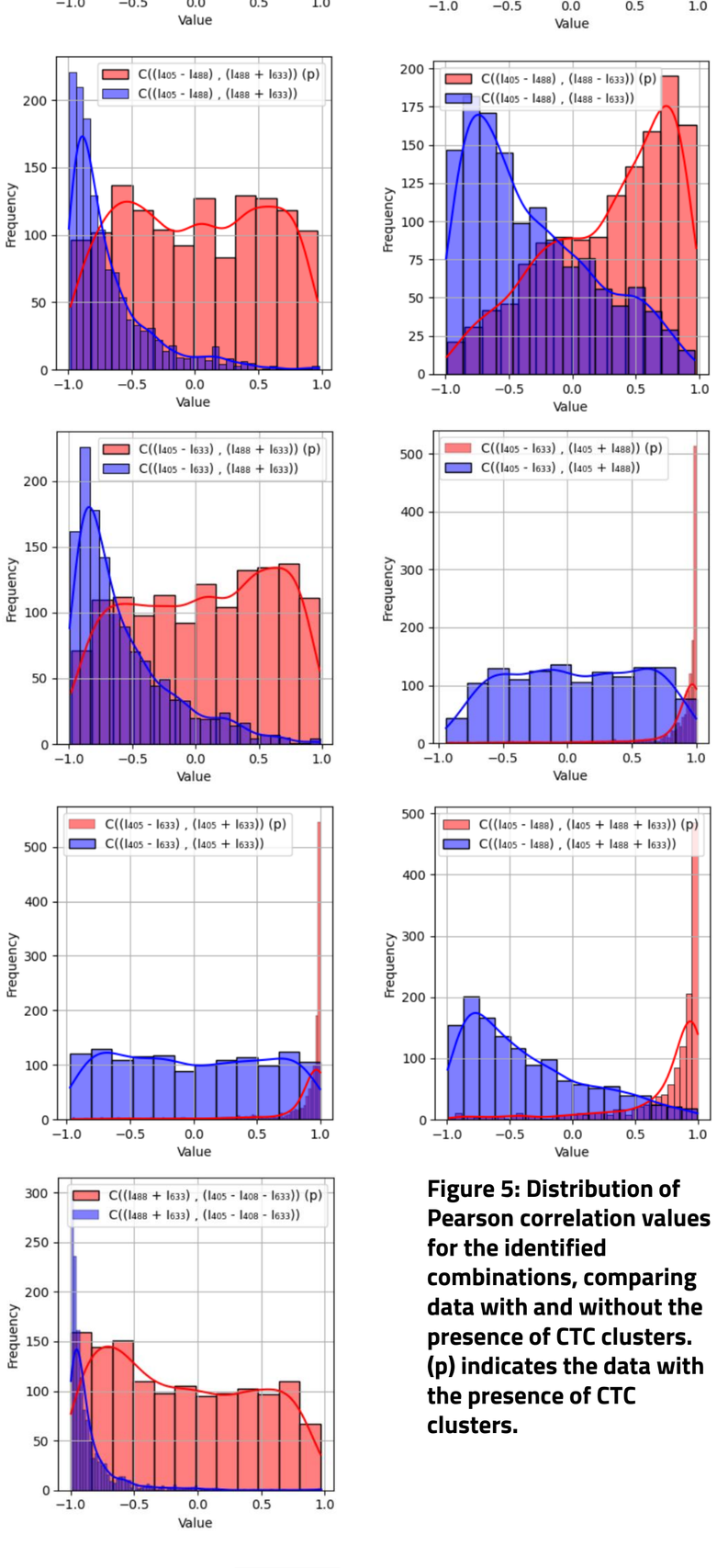


Figure 5: Distribution of Pearson correlation values for the identified combinations, comparing data with and without the presence of CTC clusters. (p) indicates the data with the presence of CTC clusters.

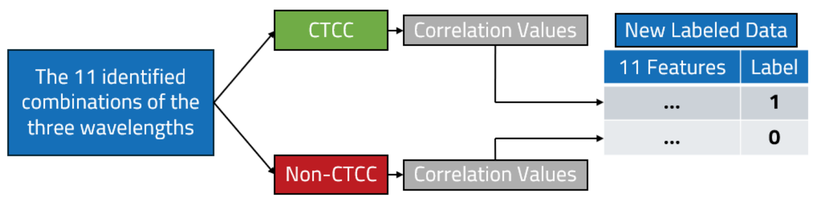


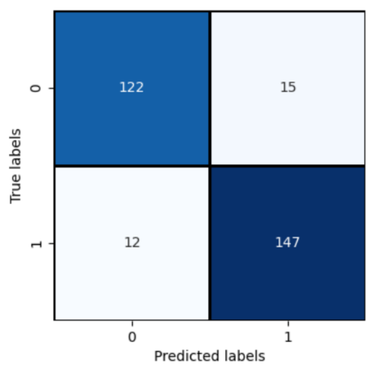
Figure 6: Flowchart depicting the correlation-based feature engineering.

Using the new labeled dataset, we trained a random forest classifier with hyperparameter tuning via a grid search method combined with stratified 5-fold cross-validation, 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 7: Confusion matrix obtained for the testing data.

FUTURE DIRECTIONS

Currently, our model is trained on segments entirely corresponding to either the presence or absence of CTC clusters, yet in practical scenarios, CTCs may be present in only a portion of a given dataset segment. For instance, the data from a single experiment comprises approximately 5.4 million data points. Within this dataset, an average of 100 to 200 data points corresponds to the presence of circulating tumor cell (CTC) clusters.

Our initial attempt to apply the current model to every 18 data points resulted in a significant number of false positives, yielding precision (purity) ranging from 2% to 22%. This highlights the necessity of a robust segmentation strategy to accurately capture CTC presence within data segments and mitigate false positives.

Therefore, future work will focus on developing improved segmentation methods and exploring training models specifically designed to address the imbalanced scenario, ensuring better precision and reliability in CTC clusters detection. This will be followed by rigorous testing to optimize performance under real-world conditions.

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