PNAS Nexus, 2024, 3, 1–8
https://doi.org/10.1093/pnasnexus/pgae048
Advance access publication 9 February 2024
Research Report

Training an automated circulating tumor cell classifier when the true classification is uncertain

Afroditi Nanou[a], Nikolas H. Stoecklein[b], Daniel Doerr[c], Christiane Driemel[b], Leon W. M. M. Terstappen[a,d] and Frank A. W. Coumans[a,d,*]

[a]Department of Medical Cell BioPhysics, Faculty of Science and Technology, University of Twente, Enschede 7522 NH, The Netherlands
[b]Department of General, Visceral and Pediatric Surgery, Heinrich-Heine University, University Hospital Düsseldorf, Düsseldorf 40225, Germany
[c]Institute for Medical Biometry and Bioinformatics, Heinrich Heine University, Düsseldorf, Germany
[d]Decisive Science, Amsterdam 1019 BB, The Netherlands
*To whom correspondence should be addressed: Email: frank@decisivescience.com
Edited By: Philip Furmanski

Abstract

Circulating tumor cell (CTC) and tumor-derived extracellular vesicle (tdEV) loads are prognostic factors of survival in patients with carcinoma. The current method of CTC enumeration relies on operator review and, unfortunately, has moderate interoperator agreement (Fleiss' kappa 0.60) due to difficulties in classifying CTC-like events. We compared operator review, ACCEPT automated image processing, and refined the output of a deep-learning algorithm to identify CTC and tdEV for the prediction of survival in patients with metastatic and nonmetastatic cancers. Operator review is only defined for CTC. Refinement was performed using automatic contrast maximization CM-CTC of events detected in cancer and in benign samples (CM-CTC). We used 418 samples from benign diseases, 6,293 from nonmetastatic breast, 2,408 from metastatic breast, and 698 from metastatic prostate cancer to train, test, optimize, and evaluate CTC and tdEV enumeration. For CTC identification, the CM-CTC performed best on metastatic/nonmetastatic breast cancer, respectively, with a hazard ratio (HR) for overall survival of 2.6/2.1 vs. 2.4/1.4 for operator CTC and 1.2/0.8 for ACCEPT-CTC. For tdEV identification, CM-tdEV performed best with an HR of 1.6/2.9 vs. 1.5/1.0 with ACCEPT-tdEV. In conclusion, contrast maximization is effective even though it does not utilize domain knowledge.

Keywords: automated classifier, label uncertainty, circulating tumor cell, tumor-derived extracellular vesicle, prognostic power

Significance Statement

Automation of human decisions in medical applications may be impeded by the uncertainty of the correct decisions. The strategy for automated labeling of training data presented here maximizes the contrast between cancer samples and benign samples. This approach is substantially less dependent on domain knowledge and thus needs to be considered when the domain knowledge is less certain. The contrast maximization approach to defining CTCs and tdEVs has proven to enhance the prediction of overall survival, demonstrating the potential of automation in medical applications.

Introduction

Identifying circulating tumor cells (CTCs) in blood samples from patients with cancer offers valuable information for clinical decision-making through a liquid biopsy approach (1). The standardization and technical validity of the technique are crucial for its routine use (2, 3). The CellSearch system was the first platform to semi-automatically detect CTCs in blood samples from patients with cancer and received clearance from the Food and Drug Administration for monitoring patients with metastatic breast, prostate, and colorectal cancers (4–6). The detection of CTCs in blood samples from patients with cancer using the CellSearch system has been extensively clinically validated. CTCs identified by this platform have been shown to have strong prognostic significance in a wide range of metastatic and nonmetastatic cancers, as demonstrated in thousands of patients (4–12). Several studies have also indicated the clinical utility of CellSearch CTCs for making systemic treatment decisions, especially in breast and prostate cancers (13–16). Additionally, recent imaging analysis techniques (ACCEPT) applied to processed CellSearch samples for CTC detection have revealed the presence of small cytokeratin (CK)-positive objects, known as tumor-derived extracellular vesicles (tdEVs), which have been similarly linked to poor clinical outcomes (17, 18).

Competing Interest: A.N., F.A.W.C., and L.W.M.M.T. receive research funding from Menarini Silicon Biosystems. L.W.M.M.T. and F.A.W.C. are listed as inventors on several US patents related to the CellSearch system, the rights of which are assigned to Menarini Silicon Biosystems. The rest of the authors declare no competing interest.
Received: July 18, 2023. **Accepted:** January 17, 2024

However, a persistent issue hindering the widespread adoption of CellSearch as a routine clinical tool is the high cost, time-consuming nature, and potential for error in the identification and counting of CTCs and tdEVs by trained operators, despite a high level of standardization achieved for the CellSearch system. In a comparison between 15 reviewers, the interreviewer agreement on CTCs was moderate (Fleiss' kappa 0.6) (19). About 40% of CTC can be identified with high confidence, but the rest is less certain. There is a field-wide ambition to replace cost- and time-intensive expert work with an automated workflow free of intra- and interreviewer bias and variability, as evidenced by other automated CTC classification approaches (20). Furthermore, specifying what to include as CTC is challenging due to the high-dimensional inclusion boundary, which has fluent transitions in several dimensions. For example, cells are occasionally observed that have the nuclear morphology of monocytes or granulocytes, but are negative for CD45-APC and clearly positive for CK-phycoerythrin (CK-PE). In the CellSearch system, the operator is expected to review these cells as "not CTC" due to their morphology, but the decision is subjective. More generally, including more events as CTC can increase the total number of true positives in a sample but will also increase the number of false positives. Consequently, setting the inclusion boundary involves a tradeoff between precision and recall.

The first attempt at a fully operator-free detection workflow was a deep-learning (DL) convolutional neural network (21). The DL-detected CTC (DL-CTC) had a better prognostic value than the operator-reviewed CTC (OP-CTC) in patients with metastatic breast cancer. However, evaluation of the network on samples from patients with benign breast disease showed that DL-CTCs were detected in 71% of the samples, whereas in only 8% of these samples, OP-CTCs were detected. A major contributor to this performance was that the dataset used for training and testing the DL network contained relatively easy events. Therefore, a re-training of the network was required using a more representative dataset with sufficient events near the decision boundary. However, such events are more difficult to label, and thus there will be a higher error rate in the training data. Label uncertainty is a common problem in image-based medical diagnostics (22, 23).

Identifying and counting CTCs are akin to finding a needle in a haystack. To address this challenge, we have created a processing pipeline that gradually reduces the haystack while retaining the majority of the needles (CTCs) until a decision can be made regarding all possible needles in a final refinement step. In this study, we evaluate three different refinement approaches and use the overall survival (OS) data of patients with metastatic breast and prostate cancers, as well as patients with nonmetastatic breast cancer before and after therapy, to determine the optimal specification of CTCs and tdEVs with the aim of incorporating it into the clinical setting.

Results

DL classification of DL-CTC and DL-tdEV

The test data of 12,144 events from 258 samples contained 1,809 (15%) events labeled as CTC and 1,386 (11%) events labeled as tdEV, suggesting that the strategy to enrich CTC and tdEV from the input data (supplementary material) was successful. The performance of the classifier on the test data after the last active learning iteration is shown in Table 1.

The retrained and redesigned DL classifier (supplementary material) performed substantially better (McNemar's $P = 0.01$) than the original DL classifier on the Zeune data (21). This suggests that the network changes were beneficial for performance on the "easy" Zeune dataset. For the other datasets, the CTC precision and recall in the newly labeled data were markedly lower, with the lowest precision and recall of 81 and 79%, respectively, for metastatic breast. A part of the breast samples was processed by more labor-intensive methods predating the CellSearch Autoprep, which probably contributed to poor performance for that dataset. In diagnostic leukapheresis (DLA) samples, CTC precision and recall of the classifier were 90 and 94%, respectively, suggesting that the new DL classifier can handle thumbnails with multiple events in close proximity. In the combined set, precision and recall were 91 and 94%, and when we check only the performance on events that the reviews were very confident about, the precision and recall are comparable with those obtained on the Zeune data. For tdEV, the new and old DL classifiers performed similarly on the Zeune data. As for CTC, performance on the other datasets was worse, with overall precision and recall of 90 and 94%, respectively.

The determined specificities indicate model performances; however, they do not directly translate to performance per sample because the test data are not a random draw from the events in a sample—rather, they are strongly enriched for CTC candidates. Furthermore, only a part of the DL-CTC was considered true CTC by operator review. In 569 IMMC38 samples, 79 million events were segmented. The network classified 99,137 of these events as DL-CTC. To determine precision, expert reviewers labeled a random draw of 10,000 DL-CTC with the CellSearch CTC review guidelines. About 36% of these 10,000 DL-CTC (3,630) were reviewed as OP-CTC. Extrapolated to all available archives in IMMC38, we estimate a total of 35,987 true CTCs, compared with a total of 27,673 CTCs in the original human reviews. This difference could in part be explained by the higher CTC recovery of the StarDist segmentation algorithm utilized in the pipeline, see Fig. 1B (24).

Refinement of CTC specifications

Taking into consideration the estimated 36% of true CTCs in the DL-CTC output, we expected that a refinement of the DL-CTC would improve prognostic performance. This refinement was implemented as a classifier trained on an unsupervised labeling where the contrast is maximized between (false-positive) DL-CTC in benign samples and DL-CTC in metastatic samples. For reference, we evaluated two more classical approaches for establishing a true CTC specification: a classifier trained with operator-labeled images and a classifier that is a manually designed decision tree (DD-CTC). These are discussed in supplementary material and compared in Fig. S8. The refinement was cursorily checked on the DL-CTC of 328 IMMC26 benign samples and of 569 IMMC38 samples. An ideal performance is a refinement that classifies 0% of DL-CTC in benign samples and ~36% of DL-CTC in metastatic prostate cancer samples.

Contrast-maximized CTC classifier

The contrast-maximized CTC classifier was trained using a sequence of t-distributed stochastic neighbors embedding (tSNE), k-nearest neighbors' (kNN) algorithm, and random forest. The resulting random forest performs inference for new samples after training. We used all DL-CTCs found in benign breast samples and metastatic prostate samples to create a low-dimensional map using tSNE. We then sampled events from the metastatic prostate samples to achieve the desired balance between metastatic and benign disease DL-CTCs. We identified regions in the

Table 1. The performance of DL classifier on test data from metastatic cancers.

		CTC		tdEV		All	N		
Model	**Dataset**	**Prec.**	**Recall**	**Prec.**	**Recall**	**Acc.**	**All**	**CTC**	**tdEV**
Prior (21)	Zeune[a]	93.4	97.3	**98.6**	**99.3**	96.4	9,049	1,334	1,966
Current	Zeune	**98.9**	**99.7**	98.4	99.2	**97.1**	1,908	354	482
	Breast	**81.4**	**79.3**	**76.3**	90.3	**74.8**	2,334	242	207
	Prostate	95.3	96.5	91.6	92.5	88.1	1,962	544	200
	Colorectal	88.7	89.1	80.7	94.0	83.1	1,976	256	151
	Lung	82.7	97.7	80.7	**85.9**	78.9	1,864	171	78
	DLA	90.1	94.2	92.9	93.3	90.5	2,100	242	268
	Overall	91.2	93.6	89.6	94.4	85.1	12,144	1,809	1,386
	High conf.	98.3	99.6	96.6	98.9	95.1	5,525	1,051	820

[a]Larger and other random draw than current model. DLA, diagnostic leukapheresis from patients with breast and lung cancer; Prec., precision; Acc., accuracy; High conf., only events labeled with high confidence by the reviewers. Maximum and minimum for each column indicated in bold font.

tSNE containing predominantly DL-CTC from metastatic patients using a kNN. These high metastatic density regions were labeled true CTC, and the rest not CTC. Lastly, we trained a random forest classifier to identify the high metastatic density regions as CM-CTC. The CM-CTC precision and recall can be tuned by adjusting the ratio of metastatic to benign DL-CTC in the input to the kNN.

From the performance results, illustrated in Fig. 2, we concluded that for metastatic prostate cancer, good performance is obtained for training sets containing 8–20% DL-CTC from metastatic patients. For 10%, we performed 5 repeats to assess the impact of sampling on the outcome and found a SD of 0.4, or 9% of the mean hazard ratio (HR). The maximum average HR was found for 20% DL-CTC in patients with metastatic prostate cancer. This classifier was used for further evaluations and is called CM-CTC henceforth. The median number of CM-CTC is 7, compared with 4 for operator-reviewed CTC, and 80% of metastatic samples have a CM-CTC count above the background level, compared with 68% for the original OP-CTC. The classifier labeled 0.5% of DL-CTC in benign samples as CM-CTC, and 43.3% of DL-CTC in metastatic prostate cancer as CM-CTC.

Overlap between CTC specifications

DL-CTC events can belong to one or more of the CM-CTC-CTC or the other tried classifiers. Figure S1 shows 10 random examples of thumbnails that belong to each combination and contain at least 2% of the total number of events. The bottom panel shows the percentage of all DL-CTCs that belong to each combination, and some of the most differentiating characteristics between the highest frequency combinations. The largest population is DL-CTCs that do not fall under any of the other refined CTC specifications (48%). Compared with the other combinations, these tend to have low CTC confidence according to the DL, low overlap between nucleus and CK, relatively low PE intensity and sharpness. DL-CTCs, which are also CTC by all other specifications (34% of DL-CTC), are overall high-confidence CTC by DL and have a wide range of CK-PE intensities and overlaps between the nuclear and CK-PE staining.

Refinement of tdEV specifications

A CM classifier was developed for tdEV in a similar way as well as a DD-tdEV classifier discussed in the supplementary material.

The CM-tdEV performance as a function of the balance between metastatic and benign DL-tdEVs in the training data is shown in Fig. 2. There is a clear optimal HR at 10% of DL-tdEV from metastatic prostate cancer samples. This classifier was taken for the next evaluations. The classifier considers 0.3% of DL-tdEV in benign samples and 10.0% of DL-tdEV in metastatic prostate cancer samples as CM-tdEV.

DL-tdEV events can be labeled DD-tdEV and/or CM-tdEV. Figure S2 shows 10 random examples for each combination. The largest population again is DL-tdEVs that do not fall into any of the other specifications (75%). Compared with the other major groups, they have a lower tdEV probability according to the DL, dimmer PE staining, and a smaller size. The sharpness in PE is low for the top three combinations. DD-tdEV contains smaller tdEV than CM-tdEV; 65% of "DD-tdEV, but not CM-tdEV" have a diameter <4 μm, while these small particles constitute only 9% of "DD-tdEV, as well as CM-tdEV."

Evaluation of CM-CTC and CM-tdEV

The CM specifications were evaluated based on the prognostic value of each specification with respect to OS, quantified by the HR. We evaluated this in metastatic breast and nonmetastatic breast cancer, both before and after treatment intervention. The results for metastatic prostate cancer are also shown, but because these samples have been used in training and model selection, the results on metastatic prostate are not considered for evaluation purposes. A summary of this comparison is shown in Fig. 3. Additional details may be found in Figs. S3 and S8, and Tables S1 and S2. CM-CTC is performant in all studies and comparable with or better than OP-CTC. In nonmetastatic breast, CM-CTC has a *P*-value at baseline and first follow-up of 0.001 and 0.004, respectively, when compared with 0.008 and 0.04, respectively, for operator-reviewed CTC. For tdEV, CM-tdEV is performant in the tested metastatic settings as well as in nonmetastatic breast. For the latter, CM-tdEV had a *P*-value at baseline and first follow-up of 0.06 and 0.001, respectively, when compared with 0.008 and 0.04, respectively, for ACCEPT-tdEV. Kaplan–Meier curves for all time points and specifications are shown in Figs. S4–S7.

The distributions of counts for the CTC and tdEV specifications are shown in Fig. S8. The separation between the distributions of benign vs. nonmetastatic breast is smaller than the separation between benign and metastatic disease.

Discussion

The true morphological specification of a CTC is unknown, due to its rarity and heterogeneity (25). In clinical samples, cells are observed that range from fully intact to nearly decayed, with a wide range of fluorescent intensities in all channels. Furthermore, we observe cells that have the morphology of a white blood cell, but are positive for CK-PE typical for a tumor

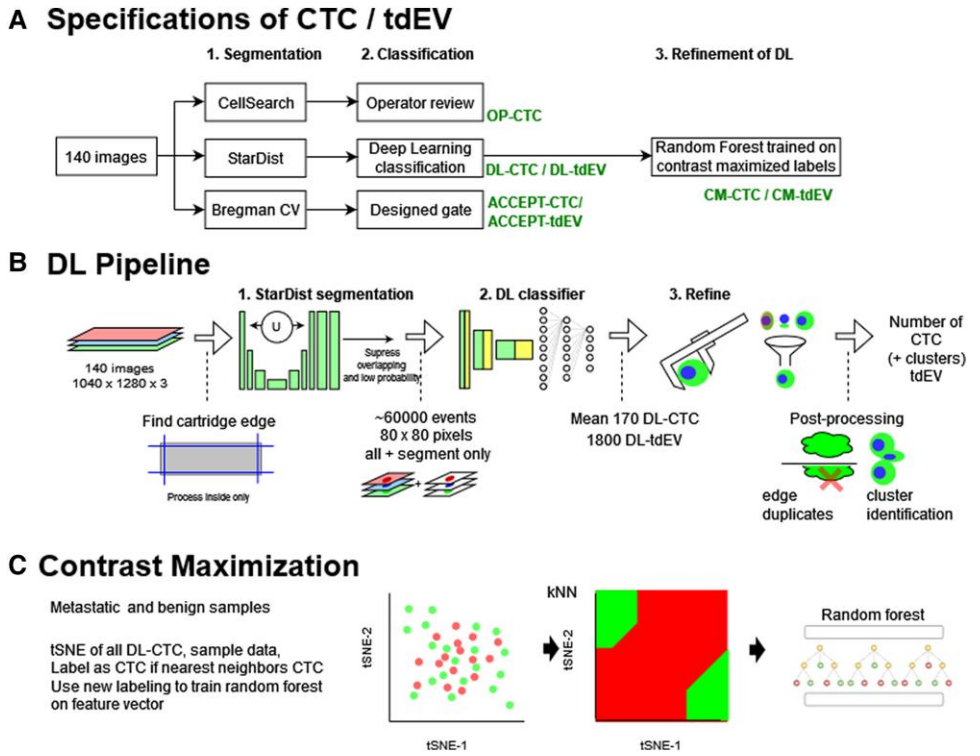

Fig. 1. Processing pipeline. The overall processing flow is shown in panel A, with the different specifications shown in green. Panel B is a graphical description of DL with refinement pipeline. Panel C is a graphical description contrast maximization approach. The dots on the tSNE plot are DL-CTC from metastatic samples in green and DL-CTC from benign samples in red. After the kNN, only events corresponding to the green area (top left and bottom right corners) are labeled as CTC for the training of a random forest classifier.

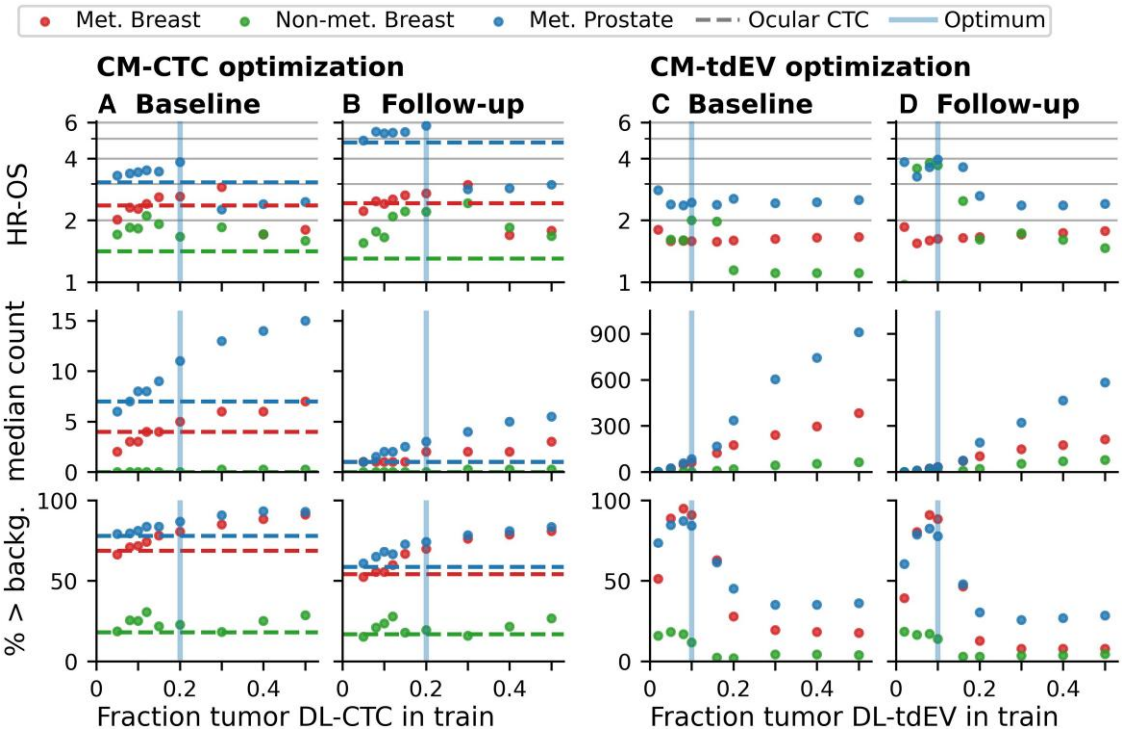

Fig. 2. The performance of CM-CTC classifiers as a function of the percentage of DL-CTC and DL-tdEV from tumor samples in the training set. The top row is the HR-OS, the middle row is the median CTC count, and the bottom row is the percentage of patient samples above the background, with background defined as the 95th percentile of samples from patients with benign breast disease. The first column is the baseline sample, taken before a therapeutic intervention, and first follow-up is the first available sample after the intervention. Dashed lines indicate the levels for the original operator-reviewed CTC. Data shown for metastatic prostate cancer and metastatic and nonmetastatic breast cancers. The bold blue vertical lines indicate the selected fraction of tumor DL-CTC (0.2, panels A and B) and tumor DL-tdEV (0.1, panels C and D) in the training that led to the optimal HR-OS.

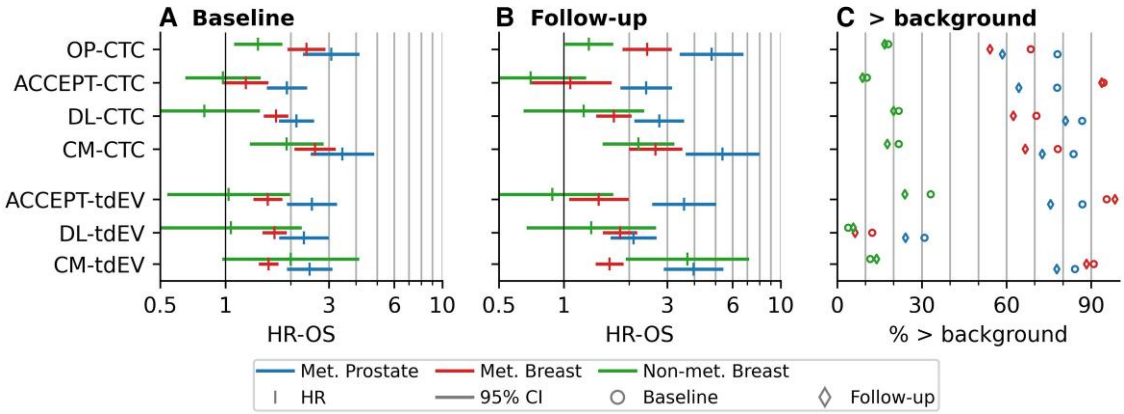

Fig. 3. Summary of HRs and percentage of samples above the background for different CTC and tdEV specifications in nonmetastatic breast, metastatic breast, and metastatic prostate cancer. HR with 95% CI of different specifications at A) baseline (before clinical intervention) and B) follow-up (after initiation of clinical intervention). HR-OS is the interquartile range HR for OS (10–90 percentile range for nonmetastatic breast due to the high survival rate). C) Percentage of samples that have a count at or above the 95th percentile of patients suspected of cancer but with benign disease.

cell. This leads to continuous transitions between normal epithelial cells, hematopoietic cells, tumor cells, and cell debris. Thus, an exact decision boundary for CTC probably does not exist, and the boundary should be viewed as a tradeoff between precision and recall. A structured evaluation of the boundary by human reviewers is not possible due to the labor needed and the poor reproducibility of human review decisions (19). We expect similar issues to exist in many image-based diagnostic tests. The refinement methods presented here do allow for a structured evaluation, but comparisons between two incremental changes are still difficult to assess due to the low number of samples and, consequently, statistically insignificant changes in the resulting HRs.

Our newly developed DL classifier for CTC enumeration surpasses the performance of the previously reported first DL classifier (21), as shown on the original benchmark test set. Because this test set contains relatively simple data, a more comprehensive data set was developed using active learning. While the old DL classifier failed in high cell density samples, we observed that the new DL classifier performed better because it had more area to classify in addition to the thumbnail that was provided to the first DL classifier. However, only 36% of DL-CTCs were also considered CTC in operator review. Some of these CTCs are simply difficult to assess, and it is possible that the DL was correct. However, the new DL network also misclassified some clear artifacts as CTCs, such as empty thumbnails or cartridge edges. For this reason, we consider the DL classifier alone to be too imprecise to replace operator review. Thus, an additional refinement step was needed. For this refinement, each DL-CTC was summarized by a set of 135 event parameters describing the shape, intensity, sharpness, overlap and similarity of fluorescent channels, and the DL class probabilities. To utilize these parameters, we needed to train another classifier using labeled data.

However, for about 40% of DL-CTCs, the true class is ambiguous. To establish true labels and train a classifier for these DL-CTCs, we applied the contrast maximization method presented here, which identifies events that are predominantly present in metastatic patients but not in patients with benign disease. It is semi-supervised, so all the available data can be utilized, and the time needed to train is just computer time. An additional advantage is that the precision–recall balance is tunable by adjusting the ratio of metastatic to benign disease events. However, it needs a larger dataset than the other approaches, and the size of this dataset is limited in most cases by the number of benign disease or healthy donor samples because these have fewer DL-CTCs. Additionally, it will identify all major differences between metastatic and benign diseases, including tdEV misclassified as DL-CTC by the DL. Lastly, a metastatic sample accidentally added to the benign data can be disastrous for performance.

Slight differences in the used fluorescence imaging systems can lead to systematic variations in image properties (26). Such variations may lead to undesired differences between parts of the dataset. For example, when greedy parameter selection was performed for the designed decision tree (supplementary material), one of the parameters describing background intensity was often among the top five parameters that could improve separation between DL-CTC from metastatic and benign samples. A classifier trained on the full set of 135 parameters would probably utilize these background parameters and would then perform poorly in new samples. See the supplementary material for a further discussion of the feature selection. Eventually, all three methods used the same 33 parameters per event, which were selected based on their use in the DD-CTCs.

CM-CTCs were equal to or better than operator CTCs to assess prognosis in patients with metastatic and nonmetastatic breast cancers. The classification model was a random forest. This is a simple model compared with some state-of-the-art machine-learning models. However, given our assumption that lack of performance primarily originates from errors in the training data, it did not seem appropriate to focus our efforts on the model side.

We tested the classifiers using OS as a measure of prognosis. However, the HR sensitivity to different CTC specifications is low. While OS can be accurately determined, other causes of death, such as adverse responses to anticancer treatment, are not related to CTC count. More robust measures, such as progression-free survival or cancer-related death, would provide a stronger link to CTC count, but they are harder to establish. Additionally, when some patients are cured after their CTC count is taken, the relationship between CTC and survival is weakened, leading to an underestimation of the link for all patients. Thus, while OS can certainly be used to evaluate a test, it cannot be used to evaluate smaller differences between CTC specifications. Furthermore, it is important to test the specifications as a continuous variable because dichotomization leads to higher uncertainty, as described elsewhere (27) and is observed from the confidence intervals for continuous HR vs. dichotomized HR in Fig. S3.

tdEV can be a complementary prognostic indicator to CTC, especially when CTC counts are low (17, 28). A possible explanation is that patient blood contains 1–2 orders of magnitude more tdEVs than CTCs (28, 29), and thus the impact of Poisson noise on the count (30) becomes negligible. However, we are not certain what should be included. The DL was trained to include all possible tdEV, including events as small as 3 pixels with very low CK-PE. As a result, DL-tdEV contained a high proportion of false positives in the opinion of a reviewer (>80%) and overall lower quality signals. The CM-tdEV made up only 10% of all DL-tdEVs but was more prognostic than ACCEPT-tdEVs (17). Although smaller tdEVs were expected to be more prognostic due to their larger numbers, larger tdEVs with brighter CK-PE staining were actually found to be more prognostic. It should be noted that the CellSearch method is not designed to detect particles <2 µm. In this case, the big advantage of CM-tdEV over other approaches is that contrast maximization does not require domain knowledge and may contribute to our insights.

In conclusion, we introduced the contract maximization method and tested it in metastatic and nonmetastatic breast cancers, where the latter presents a difficult challenge due to low CTC counts and a high risk of false-positive identifications. To effectively evaluate small variations, a more precise performance measure than OS is needed for future studies. A CM-CTC classifier is adjustable to prioritize either precision or recall. Our work represents a major advancement toward the integration of CellSearch CTCs into clinical routine. The classification pipeline presented here eliminates the need for human operators and the associated errors, time, and cost required for CTC determination.

Methods

Image datasets used for the training, test, and evaluation of the classifiers

The images used in this study are all from previously reported clinical studies using the CellSearch system. We used 418 samples from benign diseases, 6,293 from nonmetastatic breast, 2,408 from metastatic breast, and 698 from metastatic prostate cancer to train, test, optimize, and evaluate CTC and tdEV enumeration. The image datasets used specifically for the training, test, and evaluation of the classifiers are described in the supplementary material.

The study protocols of each of these studies were approved by the ethics committees of the participating institutions, and all patients signed informed consent. The CellSearch system performs immuno-magnetic enrichment of epithelial cell adhesion molecule-positive cells from blood samples and stains these cells with DAPI (nucleus), cytokeratin 8, 18, 19 conjugated to phycoerythrin (CK-PE), and lymphocyte common antigen conjugated to allophycocyanin (CD45-APC). The fluorescein isothiocyanate (FITC) channel is available for optional additional marker staining. For each sample, an archive containing ~140 four-layer immunofluorescence tiff images is stored.

Pipeline

The processing pipeline is graphically depicted in Fig. 1B and consists of: (i) segmentation of events of interest from the image datasets by StarDist, a DL-based segmentation method that has been optimized for the segmentation of cells and tdEVs from these images previously (24, 31, 32); (ii) coarse classification of these events by a DL net modified from Zeune et al. (21); (iii) further classification refinement. Until refinement, the system is optimized to retain all possible CTCs and tdEVs. For the refinement, we evaluated a kNN algorithm to find high-density regions of tumor DL-CTC in a tSNE of possible CTC in benign and tumor samples. These high-density regions are learned using a random forest for the parameters for each event.

DL classification

The segmented events of interest are classified by the DL network. This network approximately follows the design described previously (21), i.e. four convolutional layers each followed by max-pooling, and then a fully connected neural network to classify each event into one of five classes (CTC, tdEV, white blood cell, bare nucleus, and artifact). To address the poor ability of the previous DL-classifier to handle empty thumbnails and multiple events in a single thumbnail, and to obtain training data that have better coverage of actual samples, changes were made to preprocessing, network input, loss function for training, and selection of events for the training set. All changes are described in the supplementary material.

Define a true CTC

Due to the optimization of the DL for sensitivity, we expect a considerable number of false positives in the DL-CTC and tdEV counts. Therefore, we expect that a refinement of the DL output will enhance prognostic performance. This refinement will be based on features that are extracted from the thumbnails (supplementary material). Because many of the extracted features are uninformative for the classification of CTC or tdEV, and thus may introduce overfitting, only the features used for the decision tree were used for the other methods. As we are not certain on which DL-CTC should be considered true-CTC, we evaluate three possible approaches for establishing a true CTC specification.

Contrast maximization between benign and metastatic disease (CM-CTC)

The contrast maximization refinement method is weakly supervised. The only information used is whether a DL-CTC was found in a benign or in a metastatic sample. It assumes all DL-CTCs from benign are not-CTC, and that the DL-CTCs from metastatic are a mixture of true CTC and not-CTC. Here, the DL-CTCs from benign and metastatic samples were represented on a single tSNE map. Events in areas in the tSNEs that contain both DL-CTC from benign as well as DL-CTC from metastatic are assumed to be not-CTC. To identify these areas, a distance weighted kNN algorithm was applied to set the true class for each DL-CTC based on the majority class for the 49 nearest DL-CTCs. For a lower fraction of DL-CTC from metastatic samples, the kNN will set fewer regions to true CTC, selecting only those that are unique to metastatic samples. Thus, the ratio can be used to adjust precision and recall. Lastly, we trained a random forest classifier on the features of the CTC class set by the kNN algorithm. This approach requires a substantial number of benign or healthy donor samples but can be trained on a very large number of samples with minimal additional effort as it is unsupervised. The method tends to overfit if the used features have a bias between sample types. We selected the features that were also utilized in the decision tree as described in the Supplementary Methods under "Feature extraction and selection."

Code environment

Development and evaluation of the different approaches was performed in Python 3.8, utilizing Lifelines 0.25.11 (33), Matplotlib 3.3.4 (34), NumPy 1.18.5 (35), OpenCV 4.0.1 (36), Pandas 1.5.2 (37), SciPy 1.4.1 (38), Scikit-Image 0.18.1 (39), Scikit-Learn 0.24.1

(40), Sewar 0.4.4, StarDist 0.6.2 (31), Tensorflow 2.2.0 (41), and Tifffile 2021.3.31 packages and associated dependencies.

Statistical analysis

The performance of different DL models was compared using McNemar's test because the test data were sampled differently from the training data. For the contingency table of McNemar's test, white blood cells, bare nucleus, and artifact classes were recoded to "not CTC nor tdEV," so any errors between these classes are ignored.

The prognostic value of the different CTC and tdEV specifications to predict OS is used as an ultimate performance measure. The prognostic value is expressed by the HR from a Cox proportional hazards regression to the $\log_{10}$ of the CTC (or tdEV) count, where we set $\log_{10}(0)$ to −1, since 0.1 CTC/tube of blood is reasonably close to the average number of CTC in metastatic patients with 0 CTC in CellSearch (42). The HR on a continuous variable represents the change in hazard for a one unit change in the variable. Because the different specifications lead to different ranges, the continuous HRs become difficult to compare. Therefore, the shown HRs are the interquartile range HRs in metastatic settings and the 10–90 percentile range HR in nonmetastatic breast cancer. The wider range was needed in nonmetastatic breast cancer because only 58 out of 400 patients died of any cause within the follow-up period, and only 57 out of 400 progressed after surgical resection of the tumor. *P*-values were computed using the log-rank *p* test. Most of the CTC literature dichotomize patients into a "favorable" and an "unfavorable" groups. Even though this dichotomization leads to information loss (27), for comparison, we also present HR of different specifications dichotomized on the median CTC/tdEV count for metastatic samples, and dichotomized on the 90th percentile of CTC/tdEV counts for the samples of patients with nonmetastatic disease in the supplementary material. Also in the supplementary material are Kaplan–Meier curves, where we split patients into up to four groups whenever possible. The split was done on the 25th, 50th, and 75th percentiles for the metastatic setting and on the 35th, 70th, and 90th percentiles for the patients with nonmetastatic breast cancer.

Acknowledgments

The authors acknowledge Eshwari Dathathri for her help with the ACCEPT analyses.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This work was supported by research funding from Menarini Silicon Biosystems (A.N., L.W.M.M.T., F.A.W.C.), by the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A) (D.D.), and by the Multi Omics Data Science project funded from the program "Profilbildung 2020" (grant no. PROFILNRW-2020-107-A), an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia (D.D.).

Author Contributions

Conceptualization and design of the study: L.W.M.M.T., F.A.W.C., and A.N. Data collection: F.A.W.C. (selection of studies and relevant archives, splitting of data for appropriate purposes, random selection for AL labeling) and A.N. (ACCEPT processing). Analysis tools: F.A.W.C. (code, methods, established approach for inductive label from tSNE of DL-CTC), A.N. (suggestion of tSNE as basis for labeling), L.W.M.M.T. (various), and D.D. (support in clustering approach—not published, variable reduction discussions, Sobel approach for sharpness). Data analysis: F.A.W.C. (design, training, evaluation of classifiers). Labeling of data for DL training: A.N., L.W.M.M.T., F.A.W.C., N.H.S., and C.D. Writing of the paper: all authors.

Data Availability

The ∼11 Tb of image and patient data used for this publication cannot be shared because of EU-GDPR regulations. The code is owned by Menarini Silicon Biosystems, who has denied permission to disseminate the code to protect their IP rights including patentability. Summary data and code to generate the figures are shared in the Supplementary material.

References

1 Lone SN, *et al.* 2022. Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. *Mol Cancer*. 21: 1–22.
2 Neves RP, *et al.* 2021. Proficiency testing to assess technical performance for CTC-processing and detection methods in CANCER-ID. *Clin Chem*. 67:631–641.
3 Neumann MH, Bender S, Krahn T, Schlange T. 2018. ctDNA and CTCs in liquid biopsy–current status and where we need to progress. *Comput Struct Biotechnol J*. 16:190–195.
4 Cohen SJ, *et al.* 2008. Relationship of circulating tumor cells to tumor response, progression-free survival, and overall survival in patients with metastatic colorectal cancer. *J Clin Oncol*. 26: 3213–3221.
5 Cristofanilli M, *et al.* 2004. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *New Engl J Med*. 351:781–791.
6 De Bono JS, *et al.* 2008. Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. *Clin Cancer Res*. 14:6302–6309.
7 Van Dalum G, *et al.* 2015. Circulating tumor cells before and during follow-up after breast cancer surgery. *Int J Oncol*. 46:407–413.
8 Janni WJ, *et al.* 2016. Pooled analysis of the prognostic relevance of circulating tumor cells in primary breast CancerPrognostic role of CTCs in primary breast cancer. *Clin Cancer Res*. 22: 2583–2593.
9 van Dalum G, *et al.* 2015. Importance of circulating tumor cells in newly diagnosed colorectal cancer. *Int J Oncol*. 46:1361–1368.
10 Bidard F-C, *et al.* 2014. Clinical validity of circulating tumour cells in patients with metastatic breast cancer: a pooled analysis of individual patient data. *Lancet Oncol*. 15:406–414.
11 Lindsay CR, *et al.* 2019. EPAC-lung: pooled analysis of circulating tumour cells in advanced non-small cell lung cancer. *Eur J Cancer*. 117:60–68.
12 Foy V, *et al.* 2021. EPAC-lung: European pooled analysis of the prognostic value of circulating tumour cells in small cell lung cancer. *Transl Lung Cancer R*. 10:1653–1665.
13 Bidard F-C, *et al.* 2021. Efficacy of circulating tumor cell count–driven vs clinician-driven first-line therapy choice in hormone receptor–positive, ERBB2-negative metastatic breast cancer: the STIC CTC randomized clinical trial. *JAMA Oncol*. 7:34–41.

14 Fehm T, *et al.* 2010. HER2 status of circulating tumor cells in patients with metastatic breast cancer: a prospective, multicenter trial. *Breast Cancer Res Treat*. 124:403–412.

15 Bidard F-C, *et al.* 2013. Clinical application of circulating tumor cells in breast cancer: overview of the current interventional trials. *Cancer Metastasis Rev*. 32:179–188.

16 Bidard F, *et al.* 2017. Anti-HER2 therapy efficacy in HER2-negative metastatic breast cancer with HER2-amplified circulating tumor cells: results of the CirCe T-DM1 trial. *Ann Oncol*. 28:v32.

17 Nanou A, *et al.* 2020. Tumour-derived extracellular vesicles in blood of metastatic cancer patients associate with overall survival. *Br J Cancer*. 122:801–811.

18 Nanou A, *et al.* 2018. Circulating tumor cells, tumor-derived extracellular vesicles and plasma cytokeratins in castration-resistant prostate cancer patients. *Oncotarget*. 9:19283–19293.

19 Zeune LL, *et al.* 2018. How to agree on a CTC: evaluating the consensus in circulating tumor cell scoring. *Cytometry A*. 93:1202–1206.

20 Shen C, *et al.* 2023. Automatic detection of circulating tumor cells and cancer associated fibroblasts using deep learning. *Sci Rep*. 13:5708.

21 Zeune LL, *et al.* 2020. Deep learning of circulating tumour cells. *Nat Mach Intell*. 2:124–133.

22 Zhou SK, *et al.* 2021. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE Inst Electr Electron Eng*. 109:820–838.

23 Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep*. 7:17816.

24 Stevens M, *et al.* 2022. StarDist image segmentation improves circulating tumor cell detection. *Cancers (Basel)*. 14:2916.

25 Andree KC, van Dalum G, Terstappen LW. 2016. Challenges in circulating tumor cell detection by the CellSearch system. *Mol Oncol*. 10:395–407.

26 Waters JC. 2009. Accuracy and precision in quantitative fluorescence microscopy. *J Cell Biol*. 185(7):1135–1148. https://doi.org/10.1083/jcb.200903097

27 Altman DG, Royston P. 2006. The cost of dichotomising continuous variables. *BMJ*. 332:1080.

28 Nanou A, *et al.* 2023. Tumor-derived extracellular vesicles as complementary prognostic factors to circulating tumor cells in metastatic breast cancer. *JCO Precision Oncol*. 7:e2200372.

29 Coumans FAW, Doggen CJM, Attard G, de Bono JS, Terstappen LWMM. 2010. All circulating EpCAM + CK + CD45-objects predict overall survival in castration-resistant prostate cancer. *Ann Oncol*. 21:1851–1857.

30 Tibbe AG, Miller MC, Terstappen LW. 2007. Statistical considerations for enumeration of circulating tumor cells. *Cytometry A*. 71:154–162.

31 Schmidt U, Weigert M, Broaddus C, Myers G. 2018. Cell detection with star-convex polygons. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, editors. *Medical image computing and computer assisted intervention - MICCAI 2018*. Cham: Springer. p. 265–273. https://doi.org/10.1007/978-3-030-00934-2_30

32 Weigert M, Schmidt U, Haase R, Sugawara K, Myers G. 2020. Star-convex polyhedra for 3D object detection and segmentation in microscopy. In: *The IEEE Winter Conference on Applications of Computer Vision*, March 1–5, 2020, Snowmass (CO). IEEE. https://doi.org/10.1109/WACV45572.2020.9093435

33 Davidson-Pilon C. 2019. Lifelines: survival analysis in Python. *J Open Source Software*. 4:1317.

34 Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 9:90–95.

35 Harris CR, *et al.* 2020. Array programming with NumPy. *Nature*. 585:357–362.

36 Bradski G. 2000. The OpenCV library. *Dr Dobb's J of Software Tools*. 120:122–125.

37 McKinney W. 2010. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. p. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

38 Virtanen P, *et al.* 2020. Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 17:261–272.

39 van der Walt S, *et al.* 2014. Scikit-image: image processing in Python. *PeerJ*. 2:e453.

40 Pedregosa F, *et al.* 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 12:2825–2830.

41 Abadi M, *et al.* 2016 TensorFlow: a system for large-scale machine learning. arXiv 029983. https://doi.org/10.48550/arXiv.1605.08695, preprint: not peer reviewed.

42 Coumans FAW, Ligthart ST, Uhr J, Terstappen LWMM. 2012. Challenges in the enumeration and phenotyping of CTC. *Clin Cancer Res*. 18:5711–5718.