AWARD NUMBER: W81XWH-19-1-0199

TITLE: Artificial Intelligence-Based Diffraction Analysis (AIDA) for Point-of-Care Breast Cancer Classification

PRINCIPAL INVESTIGATOR: Dr. Hakho Lee

CONTRACTING ORGANIZATION: Massachusetts General Hospital, Boston, MA

REPORT DATE: July 2021

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
July 2021	Annual	01Jul2020-30Jun2021

4. TITLE AND SUBTITLE

Artificial Intelligence-Based Diffraction Analysis (AIDA) for Point-of-Care Breast Cancer Classification

5a. CONTRACT NUMBER
W81XWH-19-1-0199

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Hakho Lee, PhD

E-Mail: hlee@mgh.harvard.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MASSACHUSETTS GENERAL HOSPITAL, THE SUSAN ROUDEBUSH
5 FRUIT ST, BOSTON MA 02114-2621

AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The overall goal of this project is to advance the next generation imaging cytometer, AIDA (Artificial Intelligence-based Diffraction Analysis), for automated molecular screening on individual cancer cells. AIDA will integrate cutting-edge developments in computational optics and machine learning: digital diffraction imaging and deep neural network. First, we will implement an AIDA imaging system equipped with multiple light sources with different wavelengths. This setup will allow us to detect different molecular markers through color-based multiplexing. Nest, we will develop a deep-learning framework for cellular analyses. Specifically, we will train deep neural networks to i) recognize individual cells directly from diffraction images, ii) extract levels of molecular information, and iii) unravel hidden phenotypes for cell stratification. The combined platform will then be applied to clinical samples. Cellular samples will be obtained from breast cancer patients and will be color-stained for triple markers: HER2, ER/PR. We will then apply AIDA to image a large number of individual cells and automatically extract their features; these data will be used to construct the molecular profile of a given sample.

15. SUBJECT TERMS
breast cancer, holography, deep learning, point-of-care

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	35	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

1. INTRODUCTION

Cellular inspection using microscopy and histology remains integral for diagnosis, prognosis, and treatment decisions. The principal technique, conventional microscopy, has low throughput, requires manual inspection by trained microscopists, and often yields variable, operator-dependent results. Such drawbacks are exacerbated in resource-limited settings where pathology bottlenecks delay cancer diagnoses and potentially lead to over/under-treatment. The problem is relevant not only across low or middle-income countries, but also in the US; nearly a quarter of the US population live in rural areas, but only 10% of physicians practice in those areas8. Developing cost-effective, scalable technologies to feasibly detect (especially at early stages) and classify cancers is thus a key mandate to better manage cancer and improve survivorship. Unfortunately, no such platforms are currently available for translational testing. The ***goal*** of this proposal is to advance a new diagnostic imaging platform for on-site, high-throughput breast cancer cell screening. Termed AIDA (Artificial Intelligence Diffraction Analysis), this platform integrates cutting-edge developments in computational optics and deep learning to facilitate accurate, fast, and automated molecular analyses of breast cancer down to the single cell.

2. KEYWORDS

Breast cancer, Holography, Deep learning, Point-of-care

3. ACCOMPLISHMENTS

What were the major goals of the project?

The major goals of the first-year funding period (2020/07 - 2021/07) were two-fold.

Goal 1: Implement an AIDA platform for dual-color, single cell imaging (100% completion, Massachusetts General Hospital/ MGH).

Goal 2: Develop AIDA deep-learning framework for single cell detection and classification (90% completion, Boston Children's Hospital/ BCH).

What was accomplished under these goals?

We have made significant progresses in developing both the imaging system and deep learning algorithm.

Massachusetts General Hospital (PI, Hakho Lee)

Dual color AIDA system. We built a new, compact dual-color AIDA device. This system consisted of i) two-LEDs (λ = 540 nm, 630 nm) to image two different dyes, ii) a CMOS imager, and iii) an on-board computer for signal processing, and iv) an automatic load tray for a fluidic cartridge. A dichromatic mirror will be used to keep the light incidence from LEDs normal to an aperture (**Fig. 1**); this scheme simplified image analyses by eliminating shadow artifacts coming from different illumination angles. Diffraction images was recorded by a monochromatic 10-megapixel imager (MT9P031, On-Semiconductor) positioned underneath the sample. An on-board computer (Raspberry Pi) was programmed to control the imager, perform image analyses, and display results. The system is capable of imaging a 30 mm^2 area and up to 10^6 single cells. We will use this system to profile patient samples.

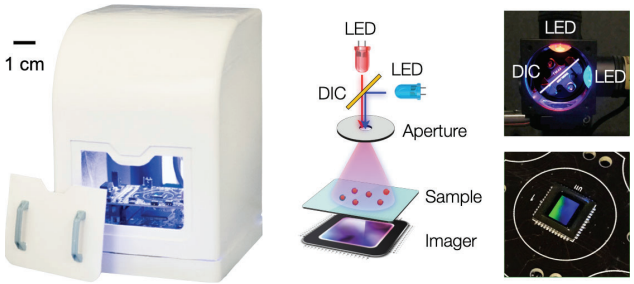


Fig. 1. AIDA imaging system. The system has a high-resolution CMOS imager and a dual-LED (λ = 470 nm, 625 nm) light source. The dichromatic mirror (DIC, 550 nm cutoff) centers light inputs to an aperture with normal incidence angle.

Chromogenic staining. We have established the overall assay protocol to color-stain cells for ER/PR (red) and HER2 (blue). We first screened 10 different chromogenic substrates reactive with horseradish peroxidase (HRP) or alkaline phosphatase (AP), and identified a pair of HRP-NovaRed and AP-Blue that showed the highest contrast and sensitivity. We next optimized the assay process. It starts by capturing cancer cells on a glass slide. These cells were then permeabilized and labeled with a cocktail of anti-ER, anti-PR, and anti-HER2 antibodies. Labeled cells were further incubated with secondary antibodies (e.g., IgG with peroxidase or alkaline phosphatase), followed by chromogens to generate color (**Fig. 2a**). For the optical detection, we found that adding an index-matching fluid (glycerol-based) was critical to reduce light scattering from cells, thereby

decreasing the intrinsic background signals (**Fig. 2b**). Applying the developed protocol, we analyzed a panel of cell lines that represents different breast cancer subtypes: MDA-MB-231, triple negative; T47D and MCF7, $(ER/PR)^{+}$ $(HER2)^{-}$; SkBr3 and BT474, $(ER/PR)^{-}(HER2)^{+}$. These cells were color-stained for ER/PR and HER2, and their diffraction patterns were obtained (**Fig. 3c**). The color intensity was proportional to the expression level of a target marker, demonstrating that AIDA can perform analytical measurements on the single cell level. The acquired images were sent to BCH site for deep-learning analyses (see below).

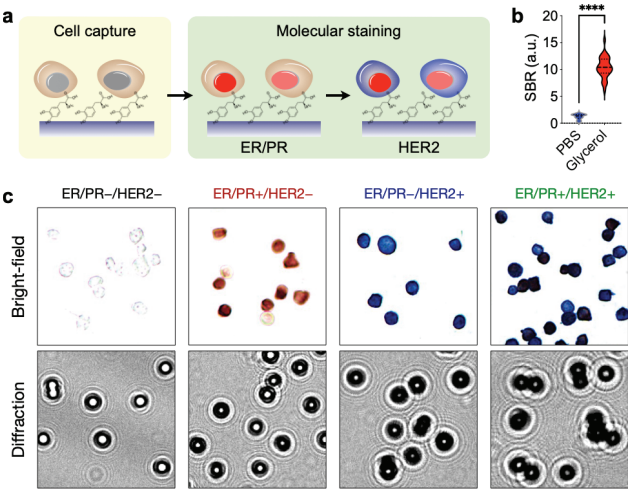


Fig. 2. Assay optimization. (a) Cancer cells were captured on a glass surface and further immobilized by applying a DOPA-based bioadhesive. Cells were then color-stained for ER/PR and HER2 expression. **(b)** Applying index-matching solution to the samples improved the signal-to-background ratio (SBR). **(c)** Quantitative molecular profiling with AIDA. Breast cancer cell lines were stained for ER/PR (red) and HER2 (blue), and imaged via bright field microscopy and AIDA.

Boston Children's Hospital (PI, Kwonmoo Lee)

Feature-fusion HoloNet for the discovery of the subtypes of breast cancer cells. In the previous reporting year, we developed a deep neural network, called HoloNet, which directly analyzes holograms for the classification of breast cancer cells based on ER/PR and HER2 intensities and the regression of the intensities. The HoloNet includes a holo-branch that extracts large features from holograms and integrate them with the small features from the standard CNN (Convolutional Neural Network). Since deep learning models can extract rich features from input images, we develop hologram feature embedding methods to identify previously uncharacterized subtypes of breast cancer cells. **Figure 3a** shows that the workflow of feature extraction and sub-clustering analysis. We developed feature-fusion HoloNet model to obtain the diffraction feature vectors. The hologram features extracted from HoloNet are used for the cell type classification and the intensity regression simultaneously. By performing multi-task learning, the HoloNet fuses the regression features with the classification features to pay more attention to the influence of the features related to the molecular intensities. Therefore, the resulting subclusters can have differential intensity distributions in each cell type. Then the feature vector is obtained from the feature fusion HoloNet model and processed by Uniform Manifold Approximation and Projection (UMAP) method to learn the feature manifold and reduce the dimension of features. **Figure 3b** represents the subcluster distribution maps with the optimal loss weight in each cell type. We obtained four subclusters with loss weight ratio 1:1 in ER/PR-HER2-, four subclusters with loss weight ratio 5:1 in ER/PR-HER2+ and ER/PR+HER2- cell types, and three subclusters with loss weight ratio 5:1 in ER/PR+HER2+.

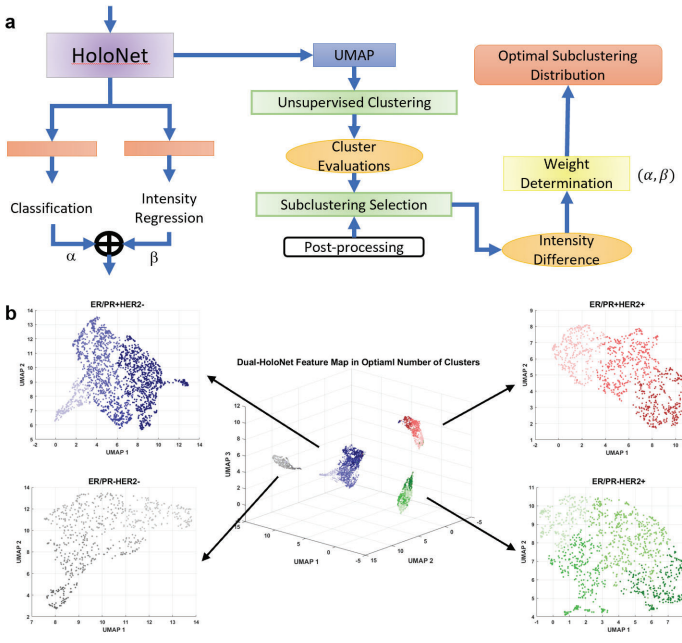


Fig. 3. Pipeline of HoloNet-based unsupervised learning. (a) The architecture of the HoloNet dual embedding model (Feature-fusion HoloNet) with unsupervised subclustering. **(b)** Distribution of hologram features from Feature-fusion HoloNet embedding and the subclustering results in each cell type.

Analysis of the identified subtypes of breast cancer cells. To know the natures of the identified subclusters of breast cancer cells, we quantified their average intensities and the distributions of the breast cancer cell lines in each subcluster. Because our hologram embedding was designed to identify the features partially discriminative to the marker intensities, the mean intensity values of the subclusters were generally statistically different (**Figs. 4a-d**). In ER/PR-HER2- cell type, the mean subcluster intensities in both channels were gradually increased from Cluster 1 to 4 (**Fig. 4a**). Most of the cells in Cluster 1 and 2 whose mean intensities

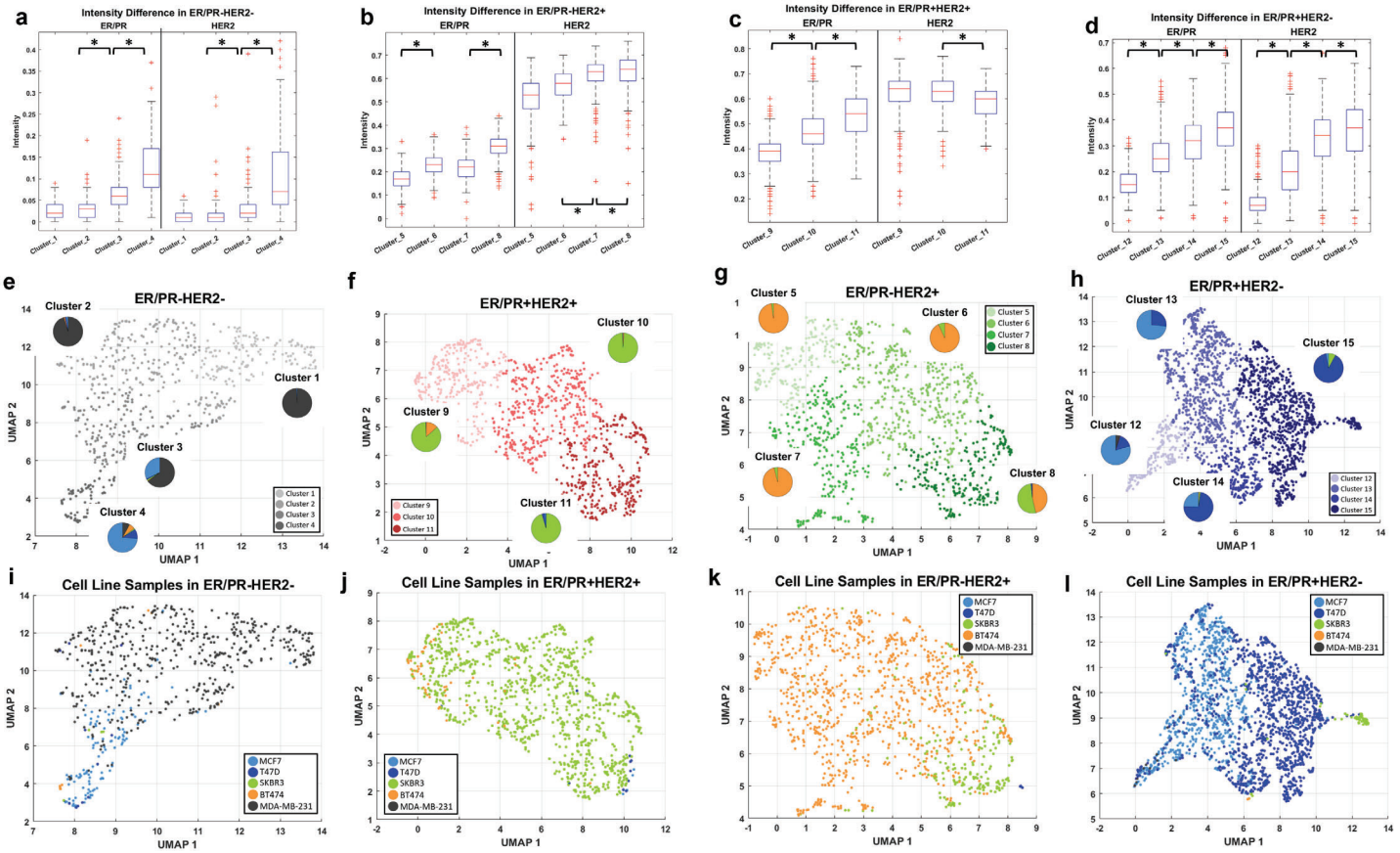


Fig. 4. Characteristics of the identified subclusters of breast cancer cells. **(a-d)** Differences of the mean intensities of the subclusters in each breast cancer cell type. **(e-h)** UMAP visualization of hologram features color-coded with the subclusters. The pie plots indicate the proportion of the cell lines in each subcluster (the color code of cell lines are in (i-l)). **(i-l)** UMAP visualization of hologram features color-coded with the cell lines. * indicates the statistical significance ($p < 0.001$).

are low were from MDA-MB-231. Cluster 3 has a mixed population of MDA-MB-231 and MCF7. Cluster 4, whose mean intensities are the highest, mainly consists of MCF7 along with minor populations from T47D, BT474, and MDA-MB-231 (**Figs. 4e, i**). In ER/PR-HER2+ cell type, the subclusters also had different mean intensities (**Fig. 4b**). Cluster 5, 6, and 7 were mainly from BT474. In Cluster 8, whose intensities are the highest among the subclusters, BT474 and SKBR3 co-existed equally (**Figs. 4f, j**). In ER/PR+HER2+ cell type, the mean intensities of ER/PR channel increased from Cluster 9 to 11 while the mean HER2 intensities decreased (**Fig. 4c**). The major cell line in these subclusters is SKBR3, but Cluster 9 has a minor cell population from BT474 (**Figs. 4g, k**). In ER/PR+HER2- cell type, the mean intensities in both channels increased from Cluster 12 to 15 (**Fig. 4d**). Cluster 12 and 13 consisted of MCF7 along with minor proportions of T47D. Cluster 14 and 15 consisted of T47D along with minor proportions of MCF7 in Cluster 14 and SKBR3 in Cluster 15 (**Figs. 4h, l**). In summary, in Cluster 3, 4, 8, 9, and 15, the cells from different cell types co-exist. Cluster 3 is near the boundary between ER/PR-HER2- and ER/PR+HER2-. Cluster 4 is near the boundary among ER/PR-HER2-, ER/PR-HER2+ and ER/PR+HER2-. Cluster 8 and 9 are near the boundary between ER/PR-HER2+ and ER/PR+HER2-. Cluster 15 is near the boundary between ER/PR+HER2- and ER/PR+HER2+. While there exist breast cancer cells whose phenotypes are near characteristics near the boundaries among the previously known cell types, these cells were treated to belong to a single cell type for a diagnostic purpose previously. We used our hologram embedding to identify those subclusters of breast cancer cells sharing similar characteristics among the known cell types.

What opportunities for training and professional development has the project provided?

Training activities.

Practical skill sets for optics and computation. [MGH] With the guidance of the PI (H. Lee) and other research fellows, the graduate student (Mr. Ismail Degani, MIT) involved in this project continues to advance his skills in GPU programming, optical systems assembly, and data acquisition. Mr. Degani is conducting this project independently. [BCH] With the guidance of the PI (Dr. Kwonmoo Lee), the research fellow and assistant (Tzu-Shi Song and Xiang Pan) involved in this project have advanced their skills on deep learning application to image datasets.

Professional development.

Course work. The concepts developed in this project (e.g., miniature optics, deep learning) has been incorporated into the intramural coursework of *CSB10 – Engineering Biosensors* taught by the PI (H. Lee), that explores key topics in biosensing.

Advanced degree. Mr. Degani (under the supervision of Dr. Hakho Lee) completed his PhD defense (July, 2021) based on the progress of this project.

Seminar. The PI (Dr. Kwonmoo Lee) presented the progress of the project at Oregon State University. Dr. Song (a research fellow in Dr. Kwonmoo Lee's lab) presented the progress of the project at Boston Children's Hospital.

How were the results disseminated to communities of interest?

Nothing to Report

What do you plan to do during the next reporting period to accomplish the goals?

[MGH site] As planned in our original Aim 3, we will perform a pilot clinical study to test AIDA's clinical utility. Cellular samples will be obtained from breast cancer patients and obtained cells will be stained for HER2 (blue), and ER/RP (red). We will then apply AIDA to i) image large number of individual cells and ii) automatically extract their molecular information. Imaging data will be sent to BCH site (Dr. K. Lee) for deep-learning analyses. The AIDA results will be compared with gold standards (e.g., histology) for concordance.

[BCH site] The clinical holograms captured in MGH site will be analyzed by our HoloNet and feature-fusion HoloNet. We will assess the accuracy of HoloNet-based marker classification and quantification. We will characterized the intra/inter-tumor heterogeneity of the clinical samples. We will confirm how these results reflect clinical pathological results. To increase the accuracy of the analysis, we will increase the resolution of hologram as follows. To restore subpixel spatial information, we will apply an EDSR deep neural network for super-resolution restoration. We will acquire the training set 1× and 4× magnified diffraction patterns from the same cells. The network will use these matching image pairs to learn spatial details. The trained network will improve the pixel density of the original diffraction patterns by 16-fold without reducing the field of view.

4. IMPACT

What was the impact on the development of the principal discipline(s) of the project?

Accurate diagnostics. Single-cell statistics contain much more diagnostic information than summary statistics. The developed imaging and computational tools will allow for automated collection of single-cell information from wide field-of-view (FOV) images. This work thus serves as the backbone for cell-based diagnostic development. Furthermore, our new staining method is compatible with intracellular markers, and by using chemically orthogonal chromogenic substrates, different colors can be generated for multiplexed detection. This approach significantly expands the power of our AIDA approach: not only the morphological data but also *molecular information* can be obtained. ***Technical breakthroughs in optical imaging.*** The digital diffraction imaging in AIDA, through lens-less image acquisition, detects $>10^4$ individual cells in a single image acquisition (no mechanical scanning). Furthermore, the diffraction patterns contain rich spatial information, which enables high resolution (~1 μm) reconstruction of original cell images. As such, the digital diffraction imaging can achieve both wide FOV and high resolving power simultaneously, overcoming the fundamental drawback of conventional microscopy. ***Powerful POC platform.*** The AIDA system is compact, user-friendly, easy-to-operate, and cost effective. Furthermore, trained neural network will make it possible to use small portable computers or handheld devices for realtime image analysis. The resulting AIDA is a self-

contained device, realizing a truly POC diagnostic system. The ability to pivot towards conventional therapies or even clinical trials matched to molecular subtypes, offers breast cancer patients powerful and timely opportunities to improve their clinical outcomes. Obviating the need for formal pathology readouts or special send outs, which can take days or weeks depending on bandwidth and resources, could accelerate the trial enrollment process and enhance personalized oncology strategies. ***Hologram embedding.*** Our hologram embedding by feature-fusion HoloNet allows us to identify the subclusters within the known cell types for refined cellular phenotyping. Some of the subclusters identified in our study have the phenotypes shared by multiple breast cancer cell types since they are located near the class boundaries in the feature space. Identifying these rare and subtle cellular phenotypes can be significant in clinical decision-making because they may have different drug sensitivity and resistance from the previously known cell types. We expect that our hologram embedding opens a new opportunity to fully characterize intra/inter-tumor heterogeneity in breast cancer and provide clinicians with valuable information for patient-specific breast cancer therapy.

What was the impact on other disciplines?
Nothing to report.

What was the impact on technology transfer?
Nothing to report.

What was the impact on society beyond science and technology?
Nothing to report.

5. CHANGES/PROBLEMS

Nothing to report.

6. PRODUCTS

Publications, conference papers, and presentations

Journal publications.

1. Liebel M, Ortega Arroyo J, Beltrán VS, Osmond J, Jo A, Lee H, Quidant R, van Hulst NF (2020) 3D tracking of extracellular vesicles by holographic fluorescence imaging. *Sci Adv* 6:eabc2508. Acknowledgement of federal support (yes).

2. Min J, Chin LK, Oh J, Landeros C, Vinegoni C, Lee J, Lee SJ, Park JY, Liu AQ, Castro CM, Lee H, Im H, Weissleder R (2020) CytoPAN-Portable cellular analyses for rapid point-of-care cancer diagnosis. *Sci Transl Med* 12:eaaz9746. Acknowledgement of federal support (yes).

3. Ortiz-Orruño U, Jo A, Lee H, van Hulst NF, Liebel M (2021) Precise Nanosizing with High Dynamic Range Holography. *Nano Lett* 21:317-322. Acknowledgement of federal support (yes).

4. Lee CY, Degani I, Cheong J, Lee JH, Choi HJ, Cheon J, Lee H (2021) Fluorescence polarization system for rapid COVID-19 diagnosis. *Biosens Bioelectron* 178:113049. Acknowledgement of federal support (yes).

5. Cheong J, Yu H, Lee CY, Lee JU, Choi HJ, Lee JH, Lee H, Cheon J (2020) Fast detection of SARS-CoV-2 RNA via the integration of plasmonic thermocycling and fluorescence detection in a portable device. *Nat Biomed Eng* 4:1159-1167. Acknowledgement of federal support (yes).

6. Park J, Park JS, Huang CH, Jo A, Cook K, Wang R, Lin HY, Van Deun J, Li H, Min J, Wang L, Yoon G, Carter BS, Balaj L, Choi GS, Castro CM, Weissleder R, Lee H (2021) An integrated magneto-electrochemical device for the rapid profiling of tumour extracellular vesicles from blood plasma. *Nat Biomed Eng* 5:678-689. Acknowledgement of federal support (yes).

7. Oh J, Carlson JCT, Landeros C, Lee H, Ferguson S, Faquin WC, Clark JR, Pittet MJ, Pai SI, Weissleder R (2021) Rapid serial immunoprofiling of the tumor immune microenvironment by fine needle sampling. *Clin Cancer Res* DOI: 10.1158/1078-0432.CCR-21-1252. Acknowledgement of federal support (yes).

Presentations
(All of the following presentations acknowledged the federal support.)

Point-of-care diagnostic technology for cancer (online)
2020 Annual Fall Meeting of the Korean BioChip Society
Hakho Lee
Jeju Island, Korea; 25-Nov-2020

Ultrafast molecular diagnostics and their application in COVID-19
Molecular Med Tri-Con Virtual Conference & Expo
Hakho Lee
Virtual meeting; 16-Feb-2021

Realizing clinical promise of liquid biopsies for cancer management
The Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy (Pittcon)
Hakho Lee
Virtual meeting; 6-Mar-2021

Deep learning-based analysis of heterogeneity of breast cancer cells using lens-free digital in-line holography
AACR AI Meeting
Kwonmoo Lee (poster presentation)
Virtual Meeting; 13-Jan-2021

Unraveling Phenotypic Heterogeneity from Live Cell Images using Deep Learning
Department of Physics, Oregon State University
Kwonmoo Lee
Virtual meeting; 1-Mar-2021

Technologies or techniques.
The research has produced a library of procedures to make optical systems, fluidic devices, and bioconjugtation (antibodies). We also advanced new neural network algorithms to rapidly analyze holograms. All data will be electronically stored and archived, and will be made available through publications in peer reviewed journals. As in the past, all of these resources will be shared freely with scientific community upon execution of a proper MTA through the Office of Corporate Licensing (MGH or BCH).

Other products
Nothing to report.

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name:	*Hakho Lee (MGH)*
Project Role:	*Principal Investigator*
Researcher Identifier	*orcid.org/0000-0002-0087-0909*
Nearest person month worked:	*2*
Contribution to Project:	*Dr. Lee supervised the overall research, interacting with investigators and research fellows, and discussing all experimental designs and data.*
Funding Support:	

Name:	*Cesar M. Castro (MGH)*
Project Role:	*Co-Investigator*
Researcher Identifier	*N/A*
Nearest person month worked:	*1*
Contribution to Project:	*Dr. Castro guided the biological research, identifying biomarkers for breast cancer detection, and validating the selection through in-vitro assays.*
Funding Support:	

Name:	*Michelle Specht (MGH)*
Project Role:	*Ci-Investigator*
Researcher Identifier	*N/A*
Nearest person month worked:	*1*
Contribution to Project:	*Dr. Specht provided translational guidance for the development of proposed imaging technology. She will help procure breast cancer specimens for diagnostic testing.*
Funding Support:	

Name:	*Ismail Degani (MGH)*
Project Role:	*Graduate student*
Researcher Identifier	*N/A*
Nearest person month worked:	*6*
Contribution to Project:	*Mr. Degani designed the holographic imaging system, constructed the deep learning framework for cell classification/segmentation, and validated the entire system.*
Funding Support:	

Name:	*Kwonmoo Lee (BCH)*
Project Role:	*Principal Investigator*
Researcher Identifier	*orcid.org/0000-0001-6838-7094*
Nearest person month worked:	*3*
Contribution to Project:	*Dr. Lee supervised the overall research, interacting with investigators and research fellows, and discussing all computational analysis results.*
Funding Support:	

Name:	*Tzu-Hsi Song (BCH)*
Project Role:	*Postdoctoral fellow*
Researcher Identifier	*N/A*
Nearest person month worked:	*10*
Contribution to Project:	*Dr.Song designed the holographic deep learning structure (Holo-Net), and perform the unsupervised learning using holograms.*
Funding Support:	

Name:	*Xiang Pan (BCH)*
Project Role:	*Research assistant*
Researcher Identifier	*N/A*

Nearest person month worked:	*8*
Contribution to Project:	*Mr. Pan optimized clustering analyses using hologram features from breast cancer cells.*
Funding Support:	

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
Nothing to report

What other organizations were involved as partners?
Nothing to report.

8. SPECIAL REPORTING REQUIREMENTS

Not applicable.

9. APPENDICES

The following manuscript uploaded in bioRxiv are attached.

Tzu Hsi Song, Mengzhi Cao, Jouha Min, Hyungsoon Im, Hakho Lee, Kwonmoo Lee (2021) Deep Learning-Based Phenotyping of Breast Cancer Cells Using Lens-free Digital In-line Holography, bioRxiv 2021.05.29.446284; doi: https://doi.org/10.1101/2021.05.29.446284. Acknowledgement of federal support (yes).

Deep Learning-Based Phenotyping of Breast Cancer Cells Using Lens-free Digital In-line Holography

Tzu Hsi Song[1,2], Mengzhi Cao[3], Jouha Min[4,5], Hyungsoon Im[4], Hakho Lee[4], Kwonmoo Lee[1,2]

[1] Department of Biomedical Engineering, Worcester Polytechnic Institute, MA, USA

[2] Vascular Biology Program and Department of Surgery, Boston Children's Hospital, Harvard. Medical School, Boston, MA, USA

[3] Data Science Program, Worcester Polytechnic Institute, Worcester, MA, USA

[4] Center for Systems Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[5] Present Address: Department of Chemical Engineering, University of Michigan Ann Arbor, MI, USA

Corresponding Authors: Kwonmoo Lee and Hakho Lee

Abstract

Lens-free digital in-line holography (LDIH) produces cellular diffraction patterns (holograms) with a large field of view that lens-based microscopes cannot offer. It is a promising diagnostic tool allowing comprehensive cellular analysis with high-throughput capability. Holograms are, however, far more complicated to discern by the human eye, and conventional computational algorithms to reconstruct images from hologram limit the throughput of hologram analysis. To efficiently and directly analyze holographic images from LDIH, we developed a novel deep learning architecture called a holographical deep learning network (HoloNet) for cellular phenotyping. The HoloNet uses holo-branches that extract large features from diffraction patterns and integrates them with small features from convolutional layers. Compared with other state-of-the-art deep learning methods, HoloNet achieved better performance for the classification and regression of the raw holograms of the breast cancer cells stained with well-known breast cancer markers, ER/PR and HER2. Moreover, we developed the HoloNet dual embedding model to extract high-level diffraction features related to breast cancer cell types and their marker intensities of ER/PR and HER2 to identify previously unknown subclusters of breast cancer cells. This

hologram embedding allowed us to identify rare and subtle subclusters of the phenotypes overlapped by multiple breast cancer cell types. We demonstrate that our HoloNet efficiently enables LDIH to perform a more detailed analysis of heterogeneity of cell phenotypes for precise breast cancer diagnosis.

Introduction

Breast cancer exhibits significant inter-tumor and intra-tumor heterogeneity, presenting significant diagnostic and therapeutic. Therefore, breast cancer tissues are biopsied to determine the hormone and growth factor receptor status for effective treatment because the hormonal status significantly affects the progress and phenotypes of breast cancer [1-3]. Specifically, the levels of nuclear estrogen (ER) or progesterone receptors (PR) are measured to determine whether breast cancer cells will respond to anti-estrogens therapy using tamoxifen, fulvestrant or aromatase inhibitors. Also, human epidermal growth factor receptor 2 (HER2) is a tyrosine kinase receptor on the surface of breast cancer cells, and HER2-positive cancers are much more likely to benefit from anti-HER2 treatment with Herceptin [1, 3]. Based on this hormone status, breast cancer can be identified into four different types: ER/PR-HER2-, ER/PR-HER2+, ER/PR+HER2+, ER/PR+HER2-. Moreover, several breast cancer subtypes have been identified in recent years, based on molecular and gene expression, and related clinical treatments are being developed [2]. Precise diagnosis and analysis of these breast cancer types, or subtypes out of heterogeneous tissue samples can provide more efficient and better treatments on breast cancer patients. However, such diagnosis is severely hampered by the limited data throughput and high cost of the current diagnosis workflow based on light microscopes.

Lens-free digital in-line holography (LDIH) has been developed to address this challenge. LDIH is a powerful imaging technique that extracts the 3D positional information of an object into a single shot of 2D interference patterns, i.e., hologram images, and computationally reconstructs a three-dimensional (3D) image of the object. In addition, LDIH has a deep observation depth that can overcome the technical limitations and extract more detailed information than conventional microscopes. LDIH has been used in various fields, such as biological sample monitoring [4-6] and cell dynamic analysis [7-8], because holographic diffraction can be obtained to extract 3D cell morphological and biochemical information in a wide field of view [3]. But there are some issues with reconstructing 3D image from a hologram image. However, reconstructing 3D images requires substantial computational resources and time, and the artifact or information loss could happen. Alternatively, if we directly use raw holograms, it is much challenging to decern meaningful features by human vision and cognition, although the diffraction patterns contain rich information.

Deep learning (DL) has recently revolutionized machine learning and is highly capable of analyzing complex, large, and high-dimensional datasets [9-11]. DL approaches can learn meaningful features within complex datasets to analyze uncharacterized image patterns and recognize hidden patterns [9-11]. Due to these advantages of feature learning, DL approaches can effectively deal with the diffraction images because the features of the diffraction patterns cannot be directly recognized or analyzed by human intuition. Moreover, DL needs a large amount of data for effective model training. Because LDIH can capture large-scale datasets with a large field of view, LDIH can help DL algorithms avoid overfitting issues and build more robust neural network models. Also, since DL approaches directly learn complex diffraction patterns, they can avoid the errors or artifacts during reconstruction processing and make the LDIH applications robust.

In recent years, DL approaches have been applied to LDIH, including reconstruction improvement [12, 13], phase retrieval [14], and classification and monitoring of various biological samples [3, 15, 16]. Min *et al.* developed an artificial intelligence diffraction analysis (AIDA) platform to make automated, rapid, high-throughput, and accurate cancer cell analysis [3]. AIDA platform allowed for quantitative molecular profiling of holograms from individual cells and revealed cellular heterogeneity. This platform can also directly perform cell recognition and color classification from raw holograms by Convolutional Neural Network (CNN) models. Moreover, Kim *et al.* have developed a deep transfer learning (DTL) approach to directly classify raw holograms generated from cells and microbeads without a reconstruction process [17]. DTL model extracts feature information using the pretrained VGG19 model as a general-purpose feature extractor [17] to identify and count microbeads on cells.

In this paper, we designed and implemented a novel deep learning approach analyze holograms from LDIH for breast cancer cell classification and analysis. **Figure 1** shows the overview of our computational framework. Firstly, we used ER/PR and HER2 immuno-stained holograms as input data. Then we developed a novel holographic deep learning architecture, termed to HoloNet learn cellular diffraction features efficiently. We used this model to identify breast cancer cell types and predict ER/PR and HER2 intensity values. Here, we demonstrated that the proposed HoloNet could efficiently extract cellular hologram features to precisely classify different cell types and estimate intensity values. After cell classification and intensity regression, a holographical deep learning network with a dual embedding model is built to learn holographic feature vectors to generate feature distribution maps. These feature distribution maps are processed by manifold learning and unsupervised clustering methods to obtain previously unknown subclusters in breast cancer cell types.

Results

Overview of Proposed Workflow

As described in **Figure 1**, the holograms were acquired by an LDIH imaging system form the breast cancer cell line (MCF7, T47D, SKBR3, BT474, and MDA-MB-231) immune-stained with anti-ER/PR and anti-HER2 conjugated with chromogens. We designed a novel holographic deep learning model to classify four breast cancer cell types: ER/PR-HER2-, ER/PR-HER2+, ER/PR+HER2+, and ER/PR+HER2- and predict intensity values of ER/PR and HER2 immuno-staining. Second, we advanced our holographical deep learning model to extract high-level features to generate feature distribution maps. Using these features, we identified previously unknown subclusters hidden in heterogeneous samples by combining manifold learning and unsupervised clustering.

Hologram Classification by HoloNet

We design a novel deep learning holographical network (HoloNet) to extract and analyze holographical features. **Figure 2(a)** shows the architecture of the proposed HoloNet model. A holo-block is built to combine local details of objects with global features using a large kernel size of the convolutional filter and a concatenated layer. The HoloNet architecture is combined with a softmax layer as the output layer to classify breast cancer cell types. The cell classification results of the proposed HoloNet model are shown in **Figure 2(b)**. Here we compare the classification performance with two types of input images, holograms and reconstructed images. We also used CNN [18-20] and Resnet [21] models to compare with the HoloNet model. First, we found that holograms provide more efficient features than reconstructed images in breast cancer cell classification. Even though the reconstructed images are built from holograms, the detailed

information may be lost during the reconstruction processing, as we mentioned before. In addition, the HoloNet model with holograms has better performance than other DL models with holograms or reconstructed images. The HoloNet model provides better accuracy and F1-score than different approaches and achieved 95.5% classification performance (**Figure 2(b)**). Although our HoloNet model has better classification performance than Densenet [22], Densenet almost provides similar performance as the HoloNet model does. It is because the HoloNet model and Densenet both used a similar concept of combining multiple-scale image features. Also, the proposed HoloNet model has better accuracy than previous work [3] over 5% increase for classifying breast cancer cell types.

Hologram Regression by HoloNet

The marker intensities of ER/PR or HER2 are the important features of phenotyping breast cancer cells. Conventionally, Therefore, we quantify their intensities directly from raw hologram using HoloNet with the fully connected layer of the regression output. The HoloNet regression model can directly and precisely obtain intensity values of ER/PR and HER2 channels without the reconstruction process. **Figure 2(c)** shows that the HoloNet model can efficiently provide the ability of intensity prediction in both ER/PR and HER2 staining channels, which the R^2 scores are 0.9743 and 0.9795, respectively. Here we also use the network structures of CNN [18-20] and Resnet [21] to predict the intensity values and compare the outcomes with our HoloNet model. We found that the HoloNet model can predict the more accurate intensities of both staining channels than other DL models.

Sub-clustering of Breast Cancer Cells with HoloNet Dual Embedding

Since deep learning models can extract rich features from input images, we further develop hologram feature embedding methods to identify previously uncharacterized subtypes of breast

cancer cells. **Figure 3(a)** shows that the workflow of feature extraction and sub-clustering analysis. We developed HoloNet dual embedding learning model, called Dual HoloNet, to obtain the diffraction feature vectors. The hologram features extracted from HoloNet are used for the cell type classification and the intensity regression simultaneously. This structure will help the HoloNet learn the features for subclustering in each breast cancer cell type while paying more attention to the features related to the intensities. Therefore, the resulting subclusters can have differential intensity distributions in each cell type. Then the feature vector is obtained from the Dual HoloNet model and processed by Uniform Manifold Approximation and Projection (UMAP) method [23] to learn the feature manifold and reduce the dimension of features. **Figure 3(b)** and **3(c)** show the feature distributions of breast cancer cells obtained by the HoloNet with holographic input images and reconstructed input images, respectively. The feature distribution map from the reconstructed input images shows that the clusters of different cell types can be distinctly separated because of the distance. But, due to the lack of intra-cluster heterogeneity, it is difficult to find potential sub-clusters. The feature distribution map from holograms exhibited high levels of intra-cluster heterogeneity while the inter-cluster distances become smaller than the features from the reconstruction images. In contrast, the hologram distribution map from the Dual HoloNet model in **Figure 3(d)** showed the sizeable intra-cluster heterogeneity and inter-cluster distances, suggesting that the features from Dual HoloNet embedding are suitable for sub-clustering analysis.

To determine the optimal number of subclusters in each cell type, we combined community detection with spectral clustering [24] and grid search. We used clustering evaluation functions to estimate the cohesion values of subclusters and used rank voting to select the optimal number of subclusters in each class. To determine the optimal balance between the branches of classification and regression, we varied the ratio of two loss weights and evaluated the sub-clustering results by

the mean intensity differences of ER/PR and HER2 among the subclusters. **Figure 3(e)** shows the mean intensity values in different ratios of the loss of classification and regression, and we chose the embedding that provided the maximum mean intensity differences.

Subclustering Analysis in Cell Line Samples

Figure 4(a) represents the subcluster distribution maps with the optimal loss weight in each cell type. We obtained four subclusters with loss weight ratio 1:1 in ER/PR-HER2-, four subclusters with loss weight ratio 5:1 in ER/PR-HER2+ and ER/PR+HER2- cell types, and three subclusters with loss weight ratio 5:1 in ER/PR+HER2+. After the subclusters were obtained in different cell type groups, we observed the cell population of subclusters in cell line sample cases. In **Figure 4(b)**, we found that MCF7 and T47D cell lines are dominated by ER/PR+HER2- cell type, but they consist of markedly different distribution of the subclusters. While Cluster 13 is the major subcluster in MCF7, Cluster 15 is the major component in T47D. Moreover, the SKBR3 cell line whose major cell type is ER/PR+HER2+ consists of Cluster 9-11 equally. In the BT474 cell line, Cluster 6 and 7 dominate ER/PR-HER2+ cell type. MDA-MD-231 cell line where most of the cells belong to ER/PR-HER2- mainly composed Cluster 1-3. In addition, Cluster 3 and 4 exist in ER/PR-HER2- of the MCF7 cell line, and Cluster 8 mainly dominates the ER/PR-HER2+ cell type of the SKBR3 cell line.

To know the natures of these subclusters, we quantified their average intensities and the distributions of the breast cancer cell lines in each subcluster. Because our hologram embedding was designed to identify the features partially discriminative to the marker intensities, the mean intensity values of the subclusters were generally statistically different (**Figure 5(a)-(d)**). In ER/PR-HER2- cell type, the mean subcluster intensities in both channels were gradually increased from Cluster 1 to 4 (**Figure 5(a)**). Most of the cells in Cluster 1 and 2 whose mean intensities are

low were from MDA-MB-231. Cluster 3 has a mixed population of MDA-MB-231 and MCF7. Cluster 4, whose mean intensities are the highest, mainly consists of MCF7 along with minor populations from T47D, BT474, and MDA-MB-231 (**Figure 5(e)** and **(i)**). In ER/PR-HER2+ cell type, the subclusters also had different mean intensities (**Figure 5(b)**). Cluster 5, 6, and 7 were mainly from BT474. In Cluster 8, whose intensities are the highest among the subclusters, BT474 and SKBR3 co-existed equally (**Figure 5(f)** and **(j)**). In ER/PR+HER2+ cell type, the mean intensities of ER/PR channel increased from Cluster 9 to 11 while the mean HER2 intensities decreased (**Figure 5(c)**). The major cell line in these subclusters is SKBR3, but Cluster 9 has a minor cell population from BT474 (**Figure 5(g)** and **(k)**). In ER/PR+HER2- cell type, the mean intensities in both channels increased from Cluster 12 to 15 (**Figure 5(d)**). Cluster 12 and 13 consisted of MCF7 along with minor proportions of T47D. Cluster 14 and 15 consisted of T47D along with minor proportions of MCF7 in Cluster 14 and SKBR3 in Cluster 15 (**Figure 5(h)** and **(l)**).

Table 1. Summary of the distributions of breast cancer cell lines in the identified subclusters.

Cell Type	ER/PR-HER2-				ER/PR-HER2+				ER/PR+HER2+			ER/PR+HER2-			
Subcluster	1	2	**3**	**4**	5	6	7	**8**	**9**	10	11	12	13	14	**15**
MCF7			+	++								++	++	+	
T47D				+								+	+	++	++
SKBR3								+	++	++	++				+
BT474				+	++	++	++	+	+						
MDA-MB-231	++	++	++	+											

As summarized in **Table 1**, in Cluster 3, 4, 8, 9, and 15, the cells from different cell types co-exist. Cluster 3 is near the boundary between ER/PR-HER2- and ER/PR+HER2-. Cluster 4 is near the boundary among ER/PR-HER2-, ER/PR-HER2+ and ER/PR+HER2-. Cluster 8 and 9 are near the boundary between ER/PR-HER2+ and ER/PR+HER2-. Cluster 15 is near the boundary between

ER/PR+HER2- and ER/PR+HER2+. While there exist breast cancer cells whose phenotypes are near characteristics near the boundaries among the previously known cell types, these cells were treated to belong to a single cell type for a diagnostic purpose previously. We used our hologram embedding to identify those subclusters of breast cancer cells sharing similar characteristics among the known cell types.

Characterizing the Heterogeneity of Breast Cancer Cells from Patients

The previous results are based on the multiple breast cancer cell lines. To confirm the clinical significance of these subclusters, we used two the hologram data from two breast cancer patient cases used in the previous study [3]. Here we used the proposed HoloNet model of classification to identify the cell type distribution of these patient cases. **Figure 6(a)** shows the proportions of the cell types from two breast cancer patients. While the cell types of the most significant proportion are ER/PR-HER2- in Patient case 1 and ER/PR+HER2- in Patient case 2, their overall cell type distributions are marginally different. We used our hologram embedding to obtain feature vectors of cellular holograms from these patients and profile them with our subclusters of breast cancer cells. In **Figure 6(b)**, we found that Patient case 1 has much more Cluster 1 from ER/PR-HER2- than Patient case 2 while Patient case 2 has much more Cluster 4 from ER/PR-HER2- than Patient case 1. Given our finding that Cluster 4 share the phenotypes from ER/PR-HER2+ and ER/PR+HER2-. The characteristics of ER/PR-HER2- in these two patients are distinct.

For ER/PR+HER2- type, two patients have the similar largest proportions of Cluster 15, sharing the similar characteristics of ER/PR+HER2+. However, Patient case 1 has a much more proportion of Cluster 13 than Patient case 2, while Patient case 2 has much more Cluster 12. Since there are significant differences in both channels between Cluster 12 and 13, the cellular phenotypes in ER/PR+HER2- of these patients are also distinct. Taken together, we demonstrated that some of

the identified subclusters existed in the breast cancer patient samples, and the subclustering results can provide much richer information on the disease status.

Discussion

We developed the HoloNet, which can efficiently learn high-level diffraction features from the complex holograms to precisely discriminate breast cancer cell types in both supervised and unsupervised learning setting. Especially, the holo-block unit adapts different large-scale filters to collect multi-scale feature information to identify detailed local cellular features. It is because the local feature information in holograms does not correspond to a particular part of cellular images but rather the entire images. These large-scale filters applied to holograms can collect more related local cellular information. Our HoloNet can efficiently extract cell information from holograms to provide better performances of cell classification and intensity regression than other existing deep learning models.

We demonstrated that the feature embedding directly from holograms enabled us to identify detailed subclusters of breast cancer cells. We added the intensity regression into the Dual HoloNet model along with the classification of the previously known cell types. This structure was able to help the neural network pay more attention to specific hologram features to enhance the difference of the marker intensities among different cell types. Then we optimized the loss weights to maximize the differences of the marker intensities among potential subclusters. This hologram embedding allowed us to identify the subclusters within the known cell types for refined cellular phenotyping. Some of the subclusters identified in our study have the phenotypes shared by multiple breast cancer cell types since they are located near the class boundaries in the feature

space. Identifying these rare and subtle cellular phenotypes can be significant in clinical decision-making because they may have different drug sensitivity and resistance from the previously known cell types. We expect that HoloNet, in conjunction with LDIH, opens a new opportunity to fully characterize intra/inter-tumor heterogeneity in breast cancer and provide clinicians with valuable information for patient-specific breast cancer therapy.

Methods

Data collection

Breast cancer cells were captured by the surface coated by antibodies (HER2, EpCAM, EGFR, and MUC1 [3]) and bio-adhesives, and then stained by anti-ER/PR and anti-HER2 conjugated with chromogen. Then cellular holograms were obtained by a LDIH system, and the image patches containing a cell were cropped with 64X64 pixel size in both hormonal staining channels. These image data include four different cell types: ER/PR-HER2-, ER/PR-HER2+, ER/PR+HER2-, and ER/PR-HER2+, from five cell lines (MCF7, T47D, SKBR3, BT474, and MDA-MB-231) [3]. The number of all holograms was 5026. For more efficient training, data augmentation was applied to balance the size of different cell types by using random rotation and flipping. Then we separated the data into training (65%), validation (15%), and testing (20%) with 5-fold cross validation for model evaluation. The ground truth of intensity values of ER/PR and HER2 of cell images were obtained from the reconstructed images.

Holographical deep learning network (HoloNet)

We constructed the HoloNet based on the concept of convolutional neural network [18-20] by adding several large-kernel-size filters to efficiently extract hologram features. Here we designed

a block called holo-block, which combines local and global holographical features. The parameters K, L, and s in a holo-block (**Figure 2(a)**) represent the kernel size of the convolutional layer, number of a feature layer, and sliding, respectively. There are three holo-blocks with different kernel size of convolutional layers, which are 16X16, 24X24, and 32X32, respectively, and average pooling layers are used to obtain and emphasize specific feature information. The number of feature layer in these blocks was 64. The sliding numbers were set to 1, 2, and 4 for each holo-block. Moreover, a convolutional layer with batch normalization and three fully connected layers are connected to these holo-blocks to build the HoloNet model. Rectified linear unit (ReLu) was used as an activation function in the model [25]. Based on the HoloNet architecture, we implemented a DL model to classify four types of breast cancer cells by a softmax layer. Then, for the intensity regression, we constructed the HoloNet architecture with a fully connected layer to predict intensity values of ER/PR and HER2 staining channels from holograms.

HoloNet dual embedding model (Dual HoloNet)

The architecture of HoloNet dual embedding model (Dual HoloNet) includes the HoloNet model with two fully connected layers of intensity regression and the cell type classification (**Figure 3(a)**). The total loss function of the training is shown as:

$$Loss_{Total} = \alpha \times Loss_{Classification} + \beta \times Loss_{Regression}$$

,where α and β are the loss weights of classification and intensity regression for loss balancing. Here we used Brier loss [26] for the classification loss as below:

$$Loss_{classification} = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2$$

,where N is the number of observations and R means the number of categorical labels. f and o represent the predictive and true label distribution, respectively. Brier loss function is similar to mean square error loss but has the same ability of loss energy as the cross-entropy function.

Neural Network Training

We used Adam optimizer with learning rate=10^{-4} and batch size=128. The categorical cross-entropy was used as a loss function for training the HoloNet model for cell classification, and the mean square error loss function was used for HoloNet model training for the intensity regression. We set the input image size as 64×64×2 for the classification and 64×64 in each staining channel for the regression. For the HoloNet model, the maximum epoch was 100 for the classification and 500 for the regression. The pixel value of input image was normalized from 0 to 1. We also automatically reduce the learning rate by multiplying with 0.1 in every 20 epochs. For training the dual HoloNet model, we set that brier loss function as a loss function for the classification and the maximum epoch=150. We used default parameters in the Keras library, and the environment in Python was TensorFlow 1.15 with CUDA 10.0 for both HoloNet models.

Unsupervised Clustering and Subcluster Selection

We used the second last layer of the Dual HoloNet model to extract the feature vector of 500 dimensions, and the UMAP method [23] was used to reduce these 500 dimensional feature vectors to three-dimensional space. In the parameters of the UMAP method, the number of the neighborhood was set to 20 and the dimension of the space was 3. Also, the minimum distance among the observation was set to 0.1, and the string metric was the correlation to compute distance in high dimensional space. Then, the pairwise distances between hologram feature vectors were calculated to generate a similarity matrix. The threshold value of similarity was set to 0.5 to build an adjacent matrix. This adjacent matrix was used as community detection to obtain subclusters

by using spectral clustering [24]. The grid search method was used to find the optimal numbers of subclusters in each cell type. We set the number of subclusters from 1 to 10 and evaluated the clustering quality of different clustering numbers using clustering evaluation functions including silhouette coefficient [27], Dunn's index [28], Calinski-Harabasz index [29], and Davies-Bouldin index [30]. Then the optimal number of subclusters in each cell type was selected by voting the highest rank from the list of clustering values among different subclustering numbers.

Loss Weight Optimization

We used grid search to determine the optimal ratio of loss weight in each cell type for subclustering. We calculated Euclidean distances of the mean intensity values of ER/PR and HER2 channels among subclusters and then averaged those Euclidean distances. We evaluated this mean Euclidean intensity distance among the subclusters with the varying loss ratios: 1:1, 3:1, 5:1, 10:1 and 100:1 (classification : regression). Then we select the optimal weight combinations which provide the maximum intensity difference in each cell type (**Figure 3(e)**).

Data availability statement

The datasets used in the current study are available from the corresponding author on a reasonable request.

Code availability statement

The datasets used in the current study are available from the corresponding author on a reasonable request.

Acknowledgement

We thank Boston Scientific for providing us with the gift for deep learning research. This work was supported by DoD grants, W81XWH1910200 (K.L.) and W81XWH1910199 (H.L.).

Competing Interests

The author declare no competing financial or non-financial interests.

Author Information

Correspondence and requests for materials, data, and code should be addressed to K.L. (kwonmoo.lee@childrens.harvard.edu) or H.L. (hlee@mgh.harvard.edu)

Reference

1. Nounou, Mohamed I., et al. "Breast cancer: conventional diagnosis and treatment modalities and recent patents and technologies." *Breast cancer: basic and clinical research* 9 (2015): BCBCR-S29420
2. Goldhirsch A, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24(9):2206–23.
3. Min, Jouha, et al. "Computational optics enables breast cancer profiling in point-of-care settings." *ACS nano* 12.9 (2018): 9081-9090.
4. Mirsky, Simcha K., et al. "Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning." *Cytometry Part A* 91.9 (2017): 893-900.
5. Singh, Dhananjay Kumar, et al. "Label-free, high-throughput holographic screening and enumeration of tumor cells in blood." *Lab on a Chip* 17.17 (2017): 2920-2932.
6. Yi, Faliu, Inkyu Moon, and Bahram Javidi. "Cell morphology-based classification of red blood cells using holographic imaging informatics." *Biomedical optics express* 7.6 (2016): 2385-2399.
7. Katz, Joseph, and Jian Sheng. "Applications of holography in fluid mechanics and particle dynamics." *Annual Review of Fluid Mechanics* 42 (2010): 531-555
8. Memmolo, Pasquale, et al. "Recent advances in holographic 3D particle tracking." *Advances in Optics and Photonics* 7.4 (2015): 713-755
9. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444, https://doi.org/10.1038/nature14539 (2015).

10. Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural networks* 61 (2015): 85-117.

11. Rawat, Waseem, and Zenghui Wang. "Deep convolutional neural networks for image classification: A comprehensive review." *Neural computation* 29.9 (2017): 2352-2449.

12. Go, T., Lee, S., You, D. *et al.* Deep learning-based hologram generation using a white light source. *Sci Rep* **10,** 8977 (2020). https://doi.org/10.1038/s41598-020-65716-4

13. Alexander, Ronald, Brian Leahy, and Vinothan N. Manoharan. "Precise measurements in digital holographic microscopy by modeling the optical train." *Journal of Applied Physics* 128.6 (2020): 060902.

14. Wu, Y. *et al*. Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery. *Optica* **5**, 704–710 (2018).

15. Ren, Z., Xu, Z. & Lam, E. Y. End-to-end deep learning framework for digital holographic reconstruction. *Adv. Photonics* **1**, 016004 (2019).

16. Zhang, G. *et al*. Fast phase retrieval in off-axis digital holographic microscopy through deep learning. *Opt. Express* **26**, 19388–19405 (2018).

17. Kim S-J, Wang C, Zhao B, Im H, Min J, Choi NR, Castro CM, Weissleder R, Lee H, Lee K (2018) Deep transfer learning-based hologram classification for molecular diagnostics. *bioRxiv* DOI:10.1101/192559

18. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 1097–1105 (2012).

19. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. *European Conference on Computer Vision*. 818–833 (Springer) (2014).

20. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. *Computer Vision and Pattern Recognition (CVPR)*, *2014 IEEE Conference on*. 1717–1724 (IEEE) (2014).

21. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

22. Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

23. McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).

24. Van Gennip, Yves, et al. "Community detection using spectral clustering on sparse geosocial data." *SIAM Journal on Applied Mathematics* 73.1 (2013): 67-83

25. Hara, Kazuyuki, Daisuke Saito, and Hayaru Shouno. "Analysis of function of rectified linear unit used in deep learning." *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015

26. Brier, Glenn W. "Verification of forecasts expressed in terms of probability." *Monthly weather review* 78.1 (1950): 1-3

27. Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53-65

28. Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.

29. Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." *Communications in Statistics-theory and Methods* 3.1 (1974): 1-27.

30. Davies, David L., and Donald W. Bouldin. "A cluster separation measure." *IEEE transactions on pattern analysis and machine intelligence* 2 (1979): 224-227.

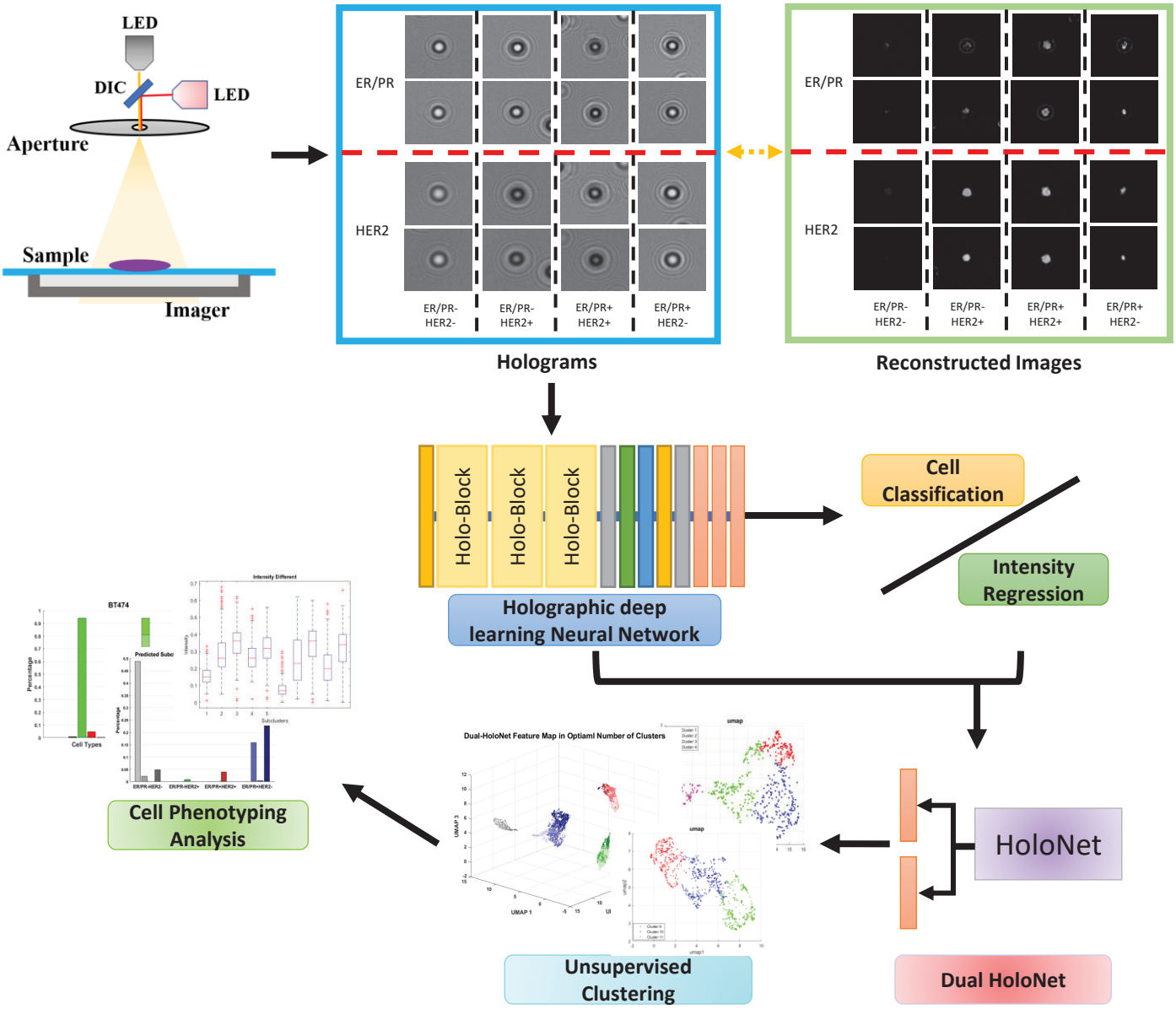


Figure 1. Workflow of the proposed deep learning models and the analysis of cell phenotyping in digital lensless inline holography.

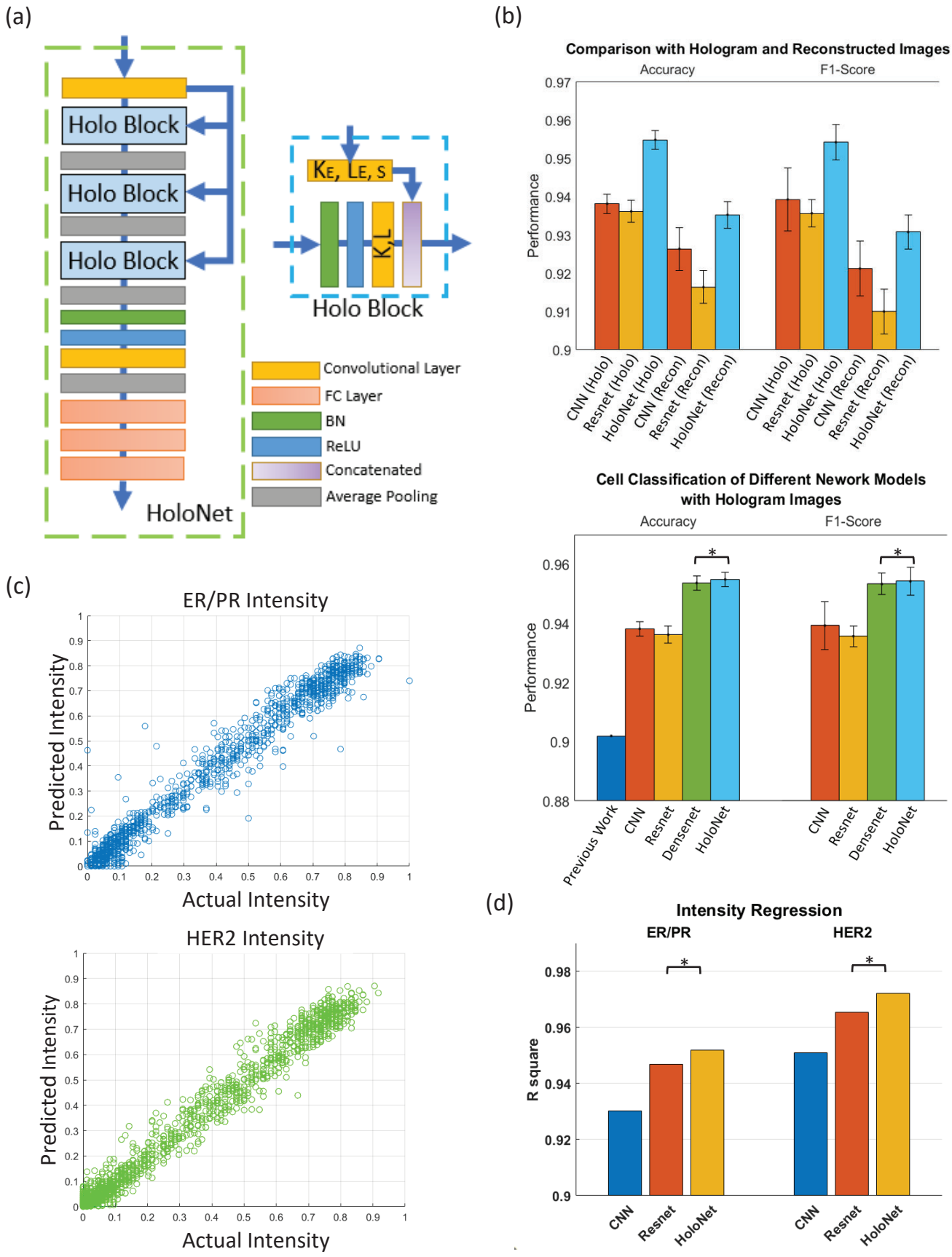


Figure 2. Structure and performance of the holographic deep learning network (HoloNet). (a) The architecture of the HoloNet. **(b)** Comparison of the performances of cell type classification between holograms and reconstructed images with different deep learning structures. **(c)** Intensity regression results from the HoloNet. **(d)** Comparison of the regression performance between different deep learning structures. * indicates the statistical significance ($p<0.01$)

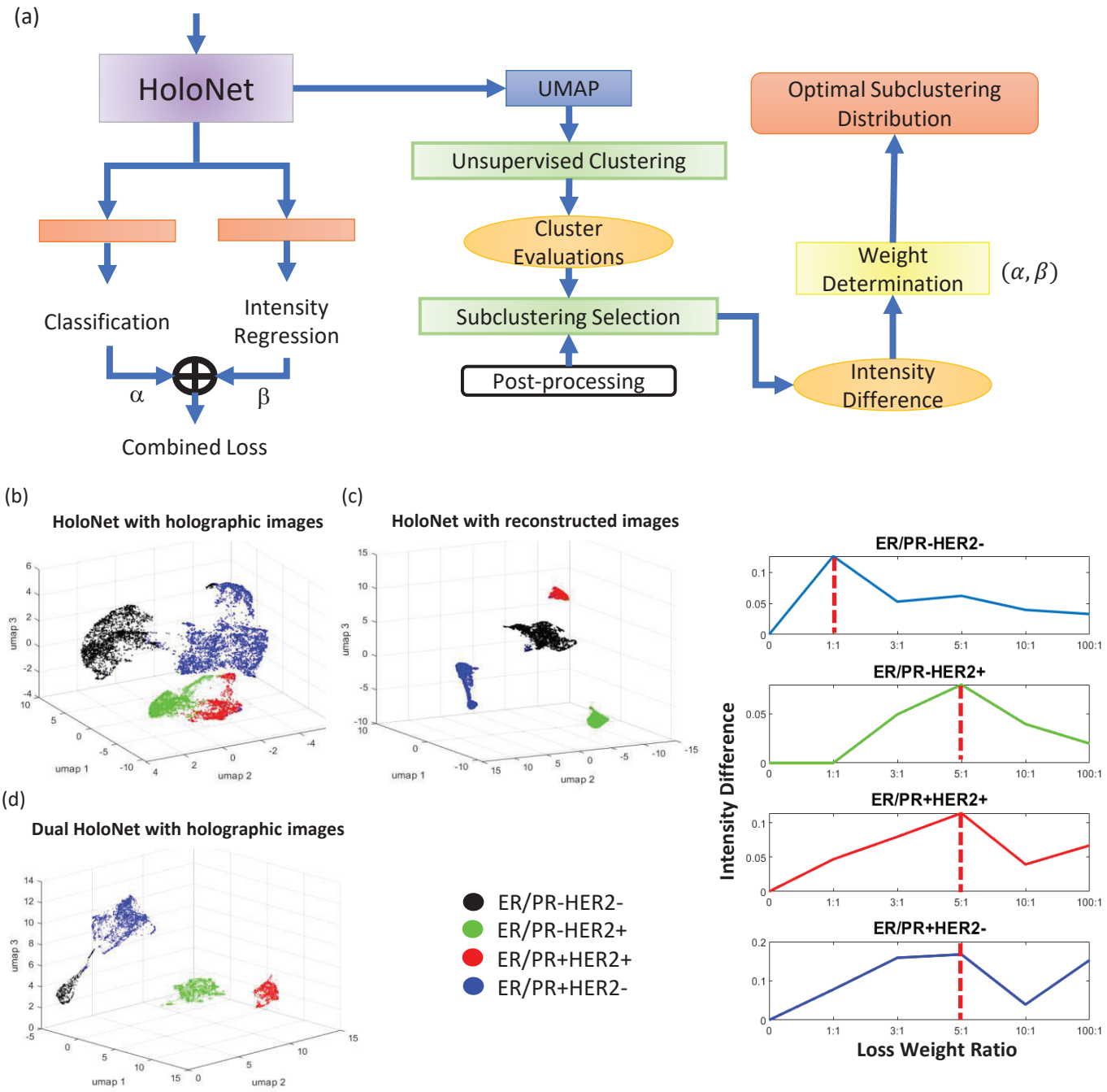
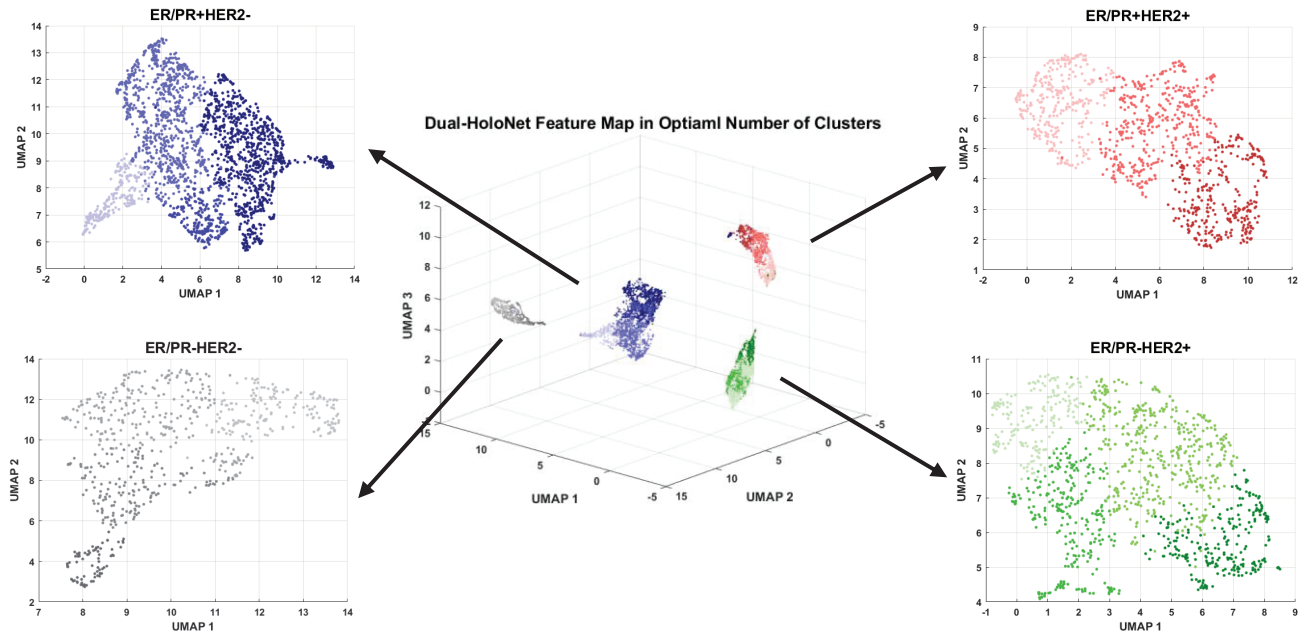


Figure 3. Pipeline of HoloNet-based unsupervised learning. (a) The architecture of the HoloNet dual embedding model (Dual HoloNet) with unsupervised subclustering selection. **(b-d)** Feature distribution maps from the HoloNet with holograms (b), the HoloNet with reconstructed images (c). the Dual HoloNet with holograms (d). **(e)** Determination of the loss weight in each cell type.

(a)



(b)

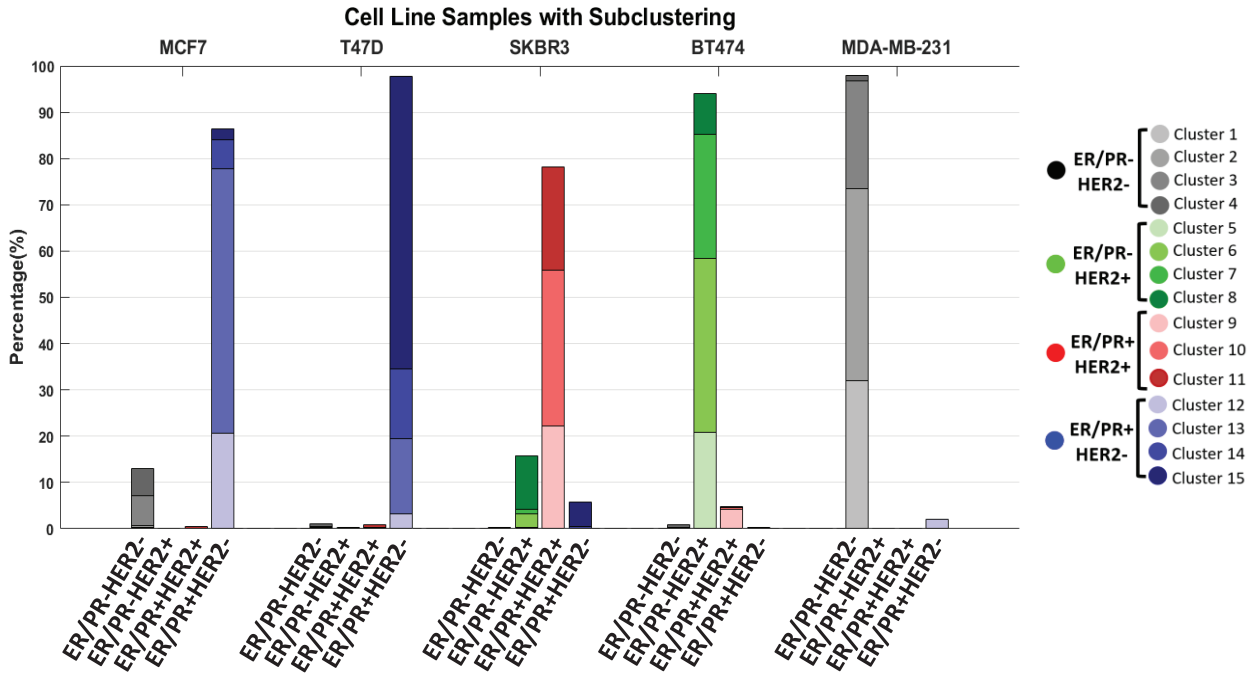


Figure 4. Subcluster results from hologram embedding. (a) Distribution of hologram features from Dual HoloNet embedding and the subclustering results in each cell type. **(b)** Subcluster distribution in each cell line.

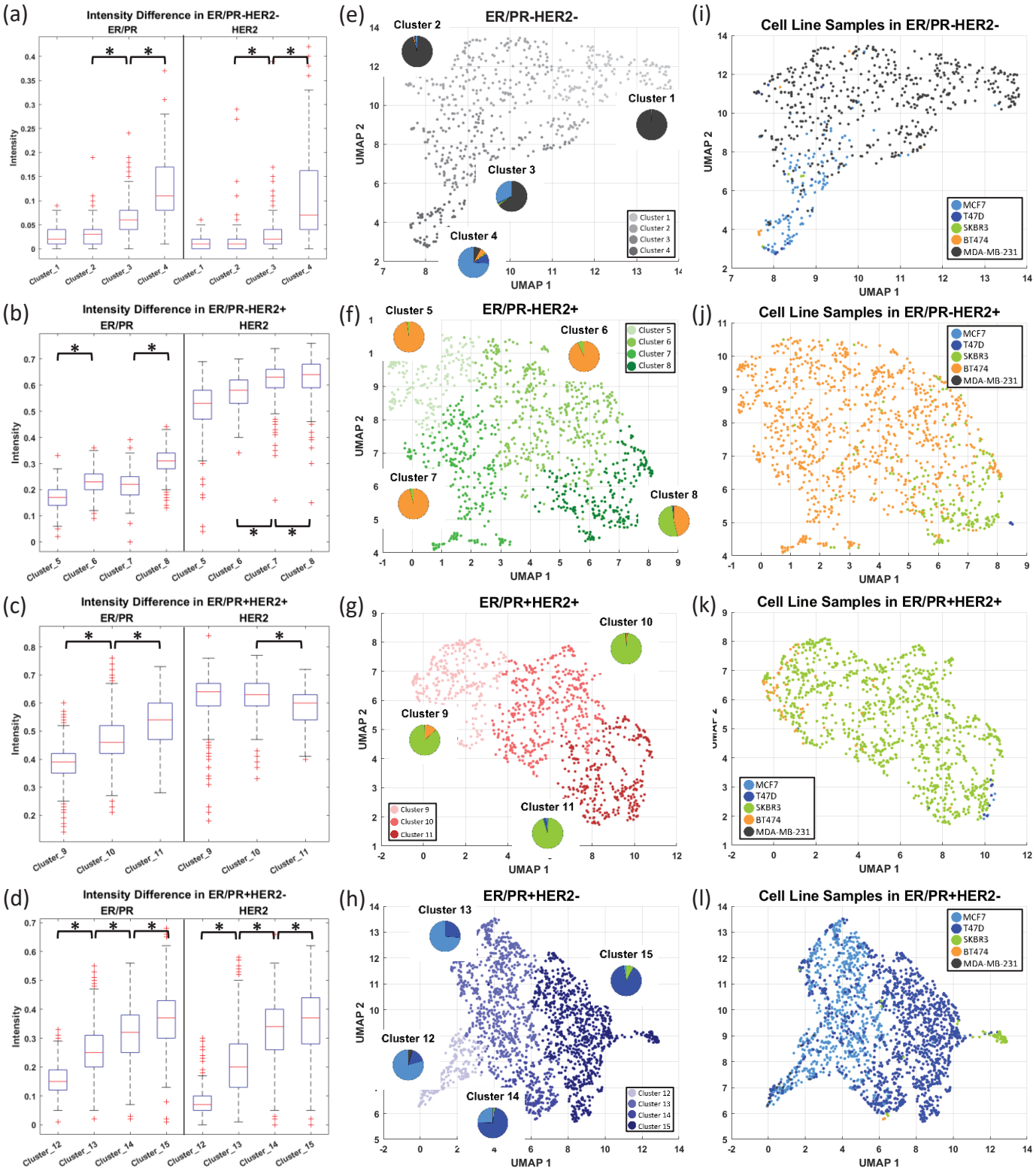


Figure 5. Characteristics of the identified subclusters of breast cancer cells. (a-d) Differences of the mean intensities of the subclusters in each breast cancer cell type. **(e-h)** UMAP visualization of hologram features color-coded with the subclusters. The pie plots indicate the proportion of the cell lines in each subcluster (the color code of cell lines are in (i-l)). **(i-l)** UMAP visualization of hologram features color-coded with the cell lines. * indicates the statistical significance ($p<0.001$)

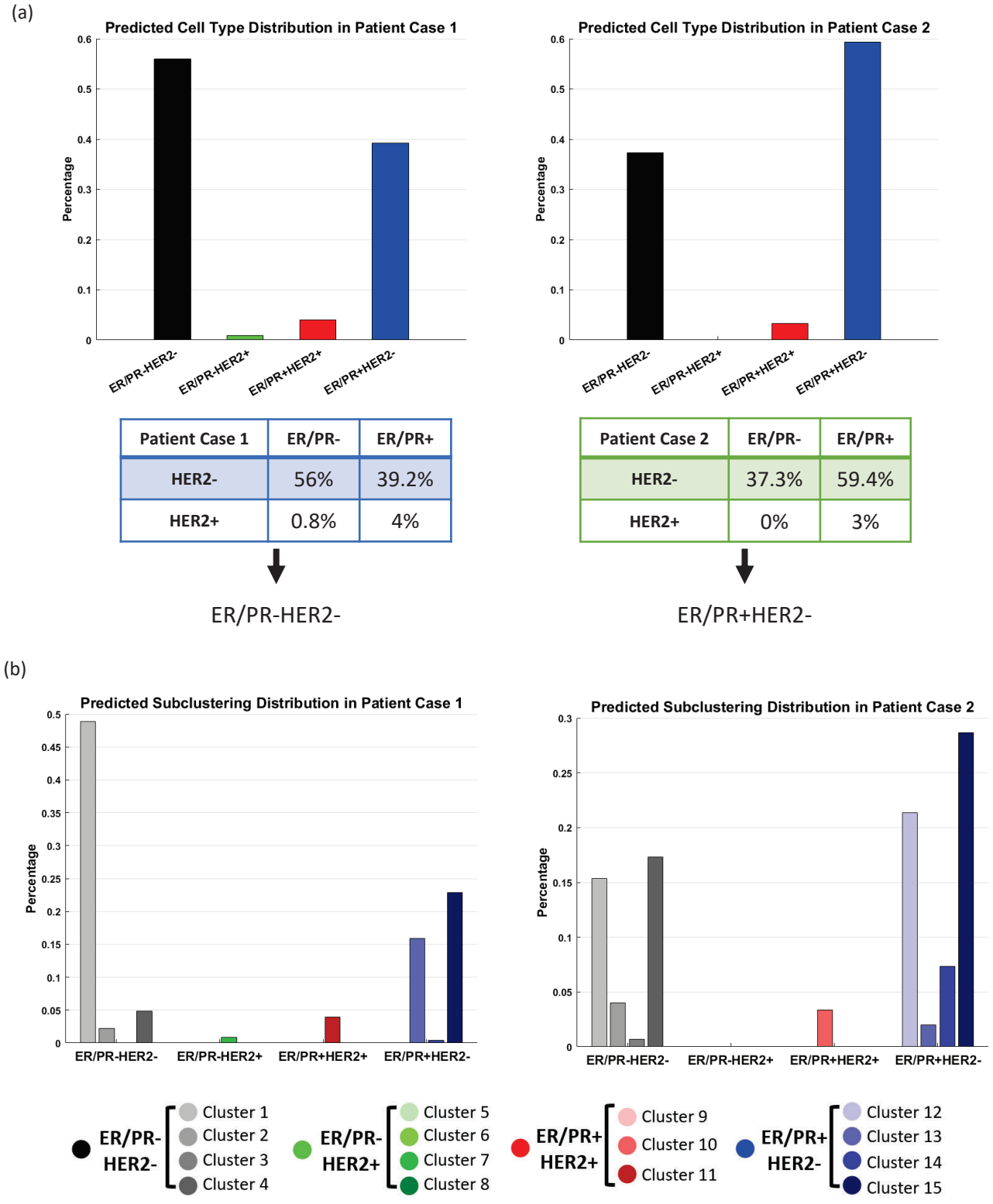


Figure 6. Profiling breast cancer cells from patient samples using the identified subclusters. (a) Distributions of the known types of breast cancer cells in two patients. **(b)** Distributions of the subclusters of breast cancer cells in two patients.