BC-MultiSet: A Multi-Target Dataset and Benchmark for Clinical Tasks in Breast Cancer

Svetlana Illarionova[1,2] s.illarionova@applied-ai.ru, Anastasiia Studenikina[1,2] st.aa73@gmail.com, Sergey Mohnenko[1,2] s.mohnenko@gmail.com, Aida Akaeva[4] aida.akaeva89@gmail.com, Elif Akbaba[3] akbabael@itu.edu.tr, Maria Boyko[1,2] maria.boyko@applied-ai.ru, Hazım Kemal Ekenel[3] ekenel@itu.edu.tr, Vyacheslav Grinevich[4] grinevich.vn@mail.ru, Rifat Hamoudi[2,5,6] rhamoudi@sharjah.ac.ae, and Maxim Sharaev[1,2,7] m.sharaev@applied-ai.ru

[1] Applied AI Institute, Moscow, Russia
[2] Biomedically Informed Artificial Intelligence Laboratory, University of Sharjah, Sharjah, United Arab Emirates
[3] Istanbul Technical University, Istanbul, Turkey
[4] National Medical Research Radiological Centre of the Ministry of Health of the Russian Federation Russia, Moscow, Russia
[5] Research Institute of Medical and Health Sciences, University of Sharjah, Sharjah, United Arab Emirates
[6] Division of Surgery and Interventional Science, University College London, London, United Kingdom
[7] Translational Oncology Research Center, QBRI, College of Health and Life Sciences, HBKU, Qatar Foundation, Doha, Qatar

Abstract. Breast cancer analysis requires integrating morphological assessment of the tumor microenvironment with molecular receptor status (ER, PR, HER2). Yet computational pathology datasets address these tasks in isolation, providing either nucleus annotations without clinical labels or molecular profiles without spatial cellular context. We present BC-MultiSet, the first histopathology dataset that bridges this gap by uniquely pairing pixel-level instance segmentation masks for three cell types (Tumor, Inflammatory, Fibroblasts) with corresponding slide-level clinical biomarkers (ER, PR, HER2). Derived from 50 invasive breast carcinoma patients, the dataset contains 5,000+ H&E patches with 109,180 nuclei manually annotated by two board-certified pathologists via consensus. We benchmark six classification architectures (including Virchow2) and three instance segmentation models (CellViT++, Cerberus, HoVer-NeXt). We also introduce multi-task learning baselines that jointly model morphology and clinical targets, revealing both the promise and challenges of integrated analysis. BC-MultiSet enables, for the first time, direct computational investigation of how cellular morphology relates to molecular expression, providing a foundational resource for developing holistic digital pathology workflows. The code is available at https://github.com/LanaLana/BC-MultiSet. The dataset is available at https://doi.org/10.5281/zenodo.20717567.

Keywords: Computational pathology · H&E histopathology · Breast cancer · Nucleus segmentation · Biomarker prediction.

1 Introduction

Therapeutic strategies for breast cancer, the most prevalent malignancy among women globally [1], are increasingly dictated by molecular profile and tumor microenvironment (TME) rather than morphology alone. Expression of estrogen (ER), progesterone (PR), and HER2 receptors defines molecular subtypes guiding clinical interventions [2]. Currently assessed via immunohistochemistry (IHC) and Fluorescence in situ hybridization (FISH), these biomarkers suffer from interpretation variability due to antibody choice, staining platforms, and pathologist subjectivity [3]. Tumor heterogeneity further complicates IHC interpretation. Meanwhile, quantifying the immune microenvironment has emerged as an independent prognostic factor, strongly correlated with patient survival and response to neoadjuvant therapy [4]. Critically, both receptor status and TME composition are initially assessed from the same hematoxylin and eosin (H&E)-stained sections, motivating computational approaches for integrated analysis. Deep learning (DL) has fundamentally transformed computational pathology by automating complex histopathological analyzes. Early convolutional neural networks, such as ResNet [5] and DenseNet [6], established the feasibility of predicting biomarkers from H&E-stained patches. More recently, Vision Transformers (ViTs) [7] have enhanced representational capacity through global self-attention mechanisms. Furthermore, pathology-specific foundation models, including Virchow2 [8], pre-trained on millions of histological images, have significantly pushed the performance ceiling for tissue classification. For instance-level tasks, methods like HoVer-Net [9] and CellViT [10] have leveraged ViT backbones to achieve state-of-the-art (SOTA) nucleus detection and classification. Despite these milestones, existing frameworks typically address biomarker prediction and cell segmentation in isolation. Integrated architectures that jointly model these tasks within a unified framework tailored for breast cancer remain underexplored.

In this work, we introduce BC-MultiSet (Breast Cancer Multi-target dataset), the first breast cancer H&E histopathology dataset providing paired per-nucleus instance segmentation annotations and slide-level molecular receptor labels for the same patient cohort. The dataset comprises 5,586 annotated patches derived from 50 WSIs of invasive breast carcinoma, with per-nucleus polygon annotations across three cell categories (Tumor, Fibroblast, Inflammatory) independently produced by two board-certified pathologists and resolved by consensus. Each patient's annotations are accompanied by a detailed clinical description, including ER, PR, HER2. The main contributions of this work are: (1) BC-MultiSet: a dataset pairing per-nucleus annotations with clinical labels from 50 patients; (2) benchmarks for six classification architectures; (3) evaluation of three segmentation models (CellViT++, HoVer-NeXt, Cerberus); (4) multi-task baselines jointly optimizing classification and segmentation.

2 BC-MultiSet Dataset

BC-MultiSet was obtained from 50 WSIs of 50 patients with invasive breast carcinoma, acquired at 40x magnification. Ethical approval was obtained from the Research Ethics Committee, and all patients provided written consent for research use. The clinical pathology reports were generated using standardized IHC and FISH assays following ASCO guidelines. The cohort comprised 35 ER-positive and 15 ER-negative cases, 27 PR-positive and 23 PR-negative cases and 16 HER2-positive and 34 HER2 negative cases. Based on the St Gallen criteria, patients were stratified into 35 luminal and 15 non-Luminal tumors, including Triple Negative and HER2-enriched subtypes. Cohort demographics: median age 56 years (IQR 45–62); histological grade I/II/III = 6/28/16; TNM stage I-II/III-IV = 33/17. Table 1 compares BC-MultiSet with existing datasets, highlighting our unique position as the only resource providing both nucleus annotations and clinical labels for the same patients. Our 50 WSIs with 109,180 annotated nuclei are consistent with established benchmarks such as MoNuSeg (30 WSIs) [11], which similarly rely on dense expert annotation of a comparatively small number of slides.

Table 1. Comparison with existing datasets.

Dataset	Patients	Patches	Clinical reports	Cell Types
MoNuSeg [11]	30	30	no	1
NuCLS [12]	∼700	∼1.7k	no	6
PanNuke [13]	–	7,901	no	5
TCGA-BRCA [14]	1,000+	—	yes	—
BC-MultiSet (Ours)	**50**	**5,586**	yes	3

A senior pathologist selected 10 regions of interest (ROIs, 1000×1000 px) per WSI, extracted blind to IHC receptor status, to ensure adequate representation of tumor tissue, stroma, and immune infiltrate.Two board-certified pathologists independently annotated all slides using the QuPath annotation tool [15]. Each pathologist delineated cell boundaries as polygon outlines (GeoJSON format) and assigned one of three biologically motivated categories: tumor cells, fibroblasts (cancer-associated fibroblasts and stromal cells) and inflammatory cells (lymphocytes, plasma cells, macrophages, neutrophils, eosinophils). Each ROI was exported as a high-resolution image tile, which was subsequently tiled into 256 x 256 pixel patches. Stain normalization was applied to all patches using the Macenko method [16], with a reference patch selected based on maximum tissue coverage and minimum background fraction. The consensus annotation process yielded 109,180 annotated nuclei across 50 slides. Median annotation time was 9 hours per WSI. Before consensus review, inter-rater agreement was substantial to excellent, with Cohen's κ of 0.83 (Tumor), 0.76 (Inflammatory), and 0.60 (Fibroblast) at IoU ≥ 0.7.

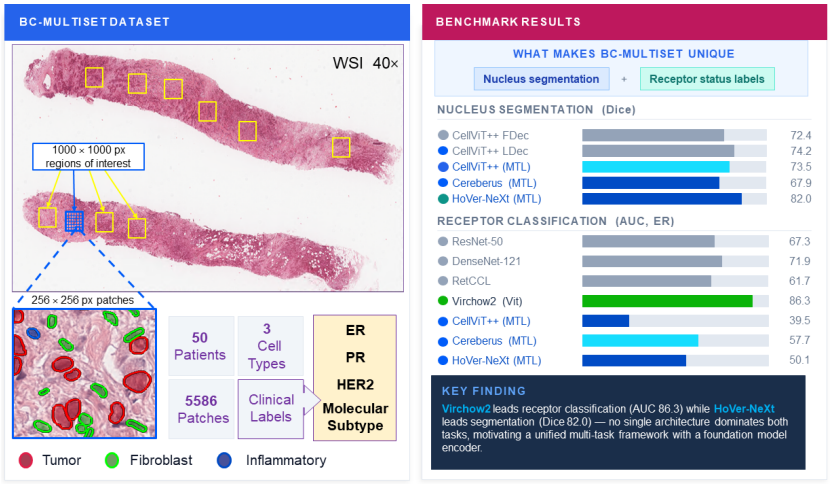

Fig. 1. A breast cancer H&E dataset pairing nucleus annotations with receptor status labels.

3 Methods

To comprehensively assess the utility of BC-MultiSet, we perform three distinct tasks: (1) patch-level classification for biomarker prediction (ER, PR, HER2, non-Luminal); (2) instance segmentation for cell type recognition (Tumor, Inflammatory, Fibroblast); (3) multi-task learning (MTL) jointly optimizing both objectives to investigate trade-offs.

For all experiments, we employed rigorous 4-fold cross-validation with patient-level splitting to ensure robust evaluation and prevent data leakage. Folds were stratified to preserve class distribution for both classification (ER, PR, HER2, non-Luminal) and segmentation tasks (Tumor, Inflammatory, Fibroblast). All patches from a given patient appear exclusively in one fold.

3.1 Classification Tasks

We benchmarked six architectures for patch-level biomarker prediction to establish baselines ranging from standard CNNs to modern foundation models. Five architectures: ResNet-50 [5], ConvNeXt-Tiny [17], DenseNet-121 (initialized with ImageNet), DenseNet-121 (pre-trained on Camelyon17), and RetCCL [18] - were trained end-to-end on H&E patches. Each model was equipped with a task-specific classification head (Dropout(0.3) + Linear $\rightarrow$ 1). Additionally, we evaluated Virchow2 [8], a ViT-H/14 pathology foundation model pre-trained on over 3 million histology images. For Virchow2, we utilized a frozen encoder approach to assess the quality of its learned representations without task-specific fine-tuning. Evaluating billion-parameter pathology foundation models with a

frozen encoder and linear probe is standard practice for small cohorts, where full fine-tuning may lead to overfitting and degrade pretrained representations.

3.2 Segmentation Tasks

For the nucleus instance segmentation task, we evaluated three SOTA architectures, all initialized with weights pre-trained on the PanNuke dataset:

1. **CellViT++** [10]: Utilizes a SAM-H encoder with a U-Net-like decoder. We evaluated two adaptation strategies: a frozen decoder (FDec) setting, where only the classification head is updated, and a learnable decoder (LDec) setting, where the decoder is fine-tuned to adapt to the specific domain of BC-MultiSet.
2. **Cerberus** [19]: A multi-branch network designed to simultaneously perform segmentation and classification tasks through shared feature extraction.
3. **HoVer-NeXt** [9]: An evolution of the HoVer-Net architecture that replaces the ResNet backbone with ConvNeXt-Tiny. This model predicts horizontal and vertical distance maps to separate clustered nuclei and is optimized for high inference speed and robust instance delineation.

3.3 Multi-Task Learning

To investigate the link between cellular morphology and clinical phenotype, we extended our segmentation architectures for joint prediction. For CellViT++, we attached lightweight binary classification heads (ER, PR, HER2, non-Luminal) to the spatially averaged SAM-H encoder features. For Cerberus, we applied global average pooling to the nuclei logits from its ResNet-34 backbone to form a compact global representation for the clinical heads. Both were trained end-to-end using a joint objective combining segmentation and classification losses: $\mathcal{L} = \mathcal{L}_{seg} + \lambda \sum_k \mathrm{CE}(y_k, \hat{y}_k)$.

Training Strategies. Naive joint training often leads to "negative transfer", where competing gradients degrade performance. To mitigate this, particularly for HoVer-NeXt, we implemented a multi-stage protocol:

1. Warmup & Segmentation: The encoder and MTL heads are initially frozen to stabilize morphological outputs via the decoders. Subsequently, the encoder is unfrozen for end-to-end segmentation training.
2. MTL Fine-tuning: Once segmentation saturates, clinical heads are introduced. We compared two strategies: **Strategy A (MTL-A)** freezes the encoder and decoders, training only the MTL heads to test the linear separability of morphological features. **Strategy B (MTL-B)** keeps the encoder frozen but fine-tunes both the segmentation decoders and MTL heads, allowing adaptation to shared representations without catastrophic forgetting.

4 Experiments and Results

4.1 Classification Tasks

To ensure a fair comparison, all CNN-based architectures were trained using identical optimization protocols and data splits. Table 2 presents the patch-level biomarker prediction results. The pathology foundation model Virchow2 achieved the strongest overall performance, yielding an AUC of 86.3 for ER and 92.3 for PR prediction. Among the end-to-end trained CNNs, ConvNeXt achieved competitive results (PR AUC 90.2), outperforming RetCCL and standard ResNets. High variance reflects the small sample size (10–13 patients per fold). Bootstrap 95% confidence intervals were wide: F1 ER classification RetCCL 43.1–77.2%, DenseNet121 36.4–75.0%; non-Luminal ResNet50 23.5–74.1%, Virchow2 21.1–62.1%, and should be interpreted cautiously. For the multi-task HoVer-NeXt models, the mean AUC across the four biomarkers was 53.8% for MTL-A and 53.5% for MTL-B. Patient-level F1 values are reported alongside AUC in Table 2; across-fold variability in F1 reflects the inherent challenge of threshold-dependent metrics on small validation folds.

Table 2. Biomarker classification results on BC-MultiSet. Results are reported as mean ± std across 4 folds. Cerberus (MTL), CellViT++ (MTL), HoVer-NeXt(A): frozen encoder + decoders, MTL head only. HoVer-NeXt(B): frozen encoder, trainable decoders + MTL head. Both initialized from Single segmentation checkpoint. Patient-level metrics throughout. **Bold** indicates best performance per metric

Model	ER		PR		HER2		non-Luminal	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Virchow2	**86.3** ± 9.8	**79.3** ± 2.4	**92.3** ± 10.2	**89.0** ± 10.6	**89.1** ± 6.6	**64.3** ± 4.8	**89.1** ± 8.0	35.1 ± 25.5
ConvNeXt	76.9 ± 7.2	78.8 ± 3.1	90.2 ± 9.8	72.1 ± 19.0	85.4 ± 12.8	63.2 ± 13.1	74.8 ± 10.0	36.7 ± 27.5
DenseNet121	71.9 ± 12.1	79.1 ± 2.2	84.5 ± 11.3	78.3 ± 4.1	79.0 ± 11.6	38.6 ± 27.0	71.0 ± 3.3	7.1 ± 14.3
ResNet50	67.3 ± 9.0	76.3 ± 4.0	80.7 ± 14.1	66.1 ± 11.2	79.5 ± 10.1	63.5 ± 13.0	72.4 ± 22.4	45.7 ± 35.8
DenseNet121-C	62.6 ± 12.8	66.8 ± 32.7	60.9 ± 19.8	59.9 ± 24.1	60.8 ± 35.4	51.7 ± 42.3	57.7 ± 14.5	25.4 ± 19.6
RetCCL	61.7 ± 19.0	53.3 ± 37.7	70.9 ± 16.3	71.0 ± 7.9	67.4 ± 29.4	13.3 ± 26.7	57.6 ± 19.2	38.1 ± 12.9
HoVer-NeXt (A)	49.7 ± 3.3	64.0 ± 37.1	53.0 ± 4.0	46.0 ± 27.2	62.6 ± 4.9	24.3 ± 25.2	49.7 ± 3.4	19.1 ± 23.2
HoVer-NeXt (B)	50.1 ± 4.4	29.9 ± 42.2	52.1 ± 5.1	51.3 ± 36.3	61.9 ± 4.7	43.5 ± 10.5	49.8 ± 4.9	33.8 ± 24.1
Cerberus (MTL)	57.7 ± 11.2	51.0 ± 8.5	56.7 ± 4.9	54.8 ± 4.3	65.2 ± 11.8	59.2 ± 9.2	57.6 ± 11.2	**50.9** ± 8.6
CellViT++ (MTL)	39.5 ± 5.6	27.6 ± 2.4	52.3 ± 7.2	33.7 ± 5.4	73.4 ± 7.7	53.8 ± 8.3	73.9 ± 8.3	48.2 ± 9.4

4.2 Segmentation Tasks

As detailed in Table 3, the single-task HoVer-NeXt model achieved strong panoptic quality ($mPQ = 40.9\%$) and CellViT++ (LDec) led in detection ($F1_{det} = 80.2\%$). This indicates that the horizontal/vertical distance map formulation combined with a ConvNeXt backbone is highly effective for separating clustered nuclei in breast cancer tissue. CellViT++ (LDec) demonstrated superior binary segmentation quality ($bPQ = 56.3\%$), outperforming HoVer-NeXt (48.7%) in

boundary delineation precision. This suggests that the SAM-H backbone provides strong shape priors, making it ideal for tasks requiring precise morphometry. The learnable decoder (LDec) configuration consistently outperformed the frozen decoder (FDec) baseline (mPQ +5.3%), confirming that adapting the segmentation head to the specific staining characteristics of BC-MultiSet is crucial. Cerberus lagged behind both modern architectures, highlighting the rapid advancement in instance segmentation methodologies.

Table 3. Segmentation and detection results on BC-MultiSet. Values are reported as mean ± std (where available). **CellViT++**: SAM-H backbone. **HoVer-NeXt**: ConvNeXt-Tiny backbone. **MTL-A**: Multi-task learning with frozen encoder and decoders (only MTL head trainable). **MTL-B**: Multi-task learning with frozen encoder (segmentation decoders and MTL head trainable). **Bold** indicates best performance.

Model	Dice↑	bPQ↑	mPQ↑	$F1_{det}$↑
Cerberus (Single)	68.2 ±2.3	25.4 ±3.8	16.2±3	34.8 ±5.2
Cerberus (MTL)	67.9 ± 2.0	24.4 ± 4.3	16.0 ± 3.5	33.6 ± 6.0
HoVer-NeXt (Single)	79.7 ±1.3	48.7 ±0.1	40.9 ±0.2	72.6 ±1.3
HoVer-NeXt (MTL-A)	80.5 ±1.2	44.6 ±0.1	37.5 ±0.3	73.8 ±1.2
HoVer-NeXt (MTL-B)	**82.0** ±0.8	46.4 ±0.1	39.6 ±0.22	73.8 ±0.8
CellViT++ (FDec)	72.4 ± 1.6	54.29 ± 0.9	36.1 ± 1.3†	78.3 ± 1.9
CellViT++ (LDec)	74.2 ± 1.3	**56.3** ± 0.7	**41.4** ± 0.9†	**80.2** ± 1.5
CellViT++ (MTL)	73.5 ±1.5	54.2 ±0.9	33.3 ± 1.4†	79.0 ±1.6

4.3 Multi-Task Learning

We analyzed the trade-off between segmentation fidelity and clinical prediction capability. For CellViT++, the multi-task setting resulted in a noticeable performance drop compared to the single-task learnable decoder (LDec) baseline (mPQ 33.3 vs. 41.4), indicating negative transfer when naively optimizing for disparate objectives. As shown in Table 3, Cerberus (MTL) experienced a slight performance drop compared to its single-task counterpart (mPQ 16.02 vs. 16.17) and lagged significantly behind the ViT- and ConvNeXt-based architectures. Cerberus' fibroblast PQ is markedly lower than that of every other model. This is due to the fact that Cerberus lacks explicit instance separation and specialized post-processing. Two HoVer-NeXt MTL strategies were evaluated: MTL-A trained only clinical heads, while MTL-B allowed decoder co-adaptation. Biomarker prediction (Table 2) remained near random for ER and non-Luminal in both strategies; HER2 showed a modest but consistent signal (AUC ≈ 62%), suggesting partial morphological discriminability. Segmentation quality was preserved at the single-task level (Table 3). HoVer-NeXt (Single) achieved the best inflammatory-cell segmentation ($PQ = 43.2$), which was better preserved by MTL-B than MTL-A, supporting decoder co-adaptation for multi-task learning.

Table 4. Class-wise segmentation results (PQ and Dice, %). Comparison of CellViT++ and HoVer-NeXt variants. **Bold** indicates best performance per metric.

Model	Tumor		Inflammatory		Fibroblasts	
	PQ ↑	Dice ↑	PQ ↑	Dice ↑	PQ ↑	Dice ↑
Cerberus (Single)	23.5 ± 4.9	57.8 ± 0.5	23.16 ± 5.1	20.8 ± 0.8	1.9 ± 0.5	0.8 ± 0.3
Cerberus (MTL)	22.6±5.5	56.6 ± 0.3	23±5.9	20.4 ± 0.6	2.0±0.5	0.9 ± 0.2
HoVer-NeXt (Single)	53.2 ± 0.5	64.5 ± 0.6	**43.2** ± 0.8	**37.6** ± 0.7	26.3 ± 0.4	**37.9** ± 0.5
HoVer-NeXt (MTL-A)	51.9 ± 0.6	**67.9** ± 0.5	36.4 ± 0.9	34.6 ± 0.8	24.3 ± 0.5	36.6 ± 0.6
HoVer-NeXt (MTL-B)	52.3 ± 0.4	67.7 ± 0.4	40.4 ± 0.7	37.4 ± 0.6	26.0 ± 0.3	37.1 ± 0.4
CellViT++ (FDec)	48.6±1.2	61.0±1.9	34.1±3.3	27.6±2.3	25.6±1.7	25.7±3.1
CellViT++ (LDec)	**53.5±1.3**	67.3±1.6	42.4±2.0	37.1±2.5	**28.2±1.4**	31.1±2.6
CellViT++ (MTL)	47.0 ± 1.5	61.0 ± 1.8	27.7 ± 3.5	29.6 ± 2.2	25.3 ± 1.8	33.0 ± 2.9

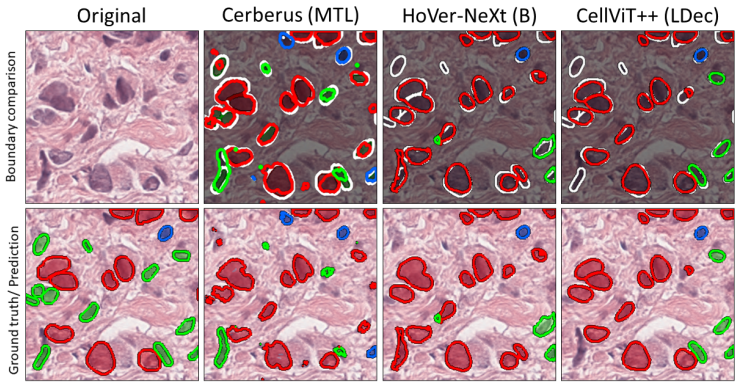

Fig. 2. Qualitative assessment of nuclei segmentation on the BC-MultiSet. The top row illustrates the boundary comparison, where **white** contours represent the ground truth and colored contours signify model predictions, red = Tumor, blue = Inflammatory, green = Fibroblast. The bottom row displays the predicted instance masks overlaid on the original H&E patch.

5 Discussion and Conclusion

The primary contribution of this work is BC-MultiSet, a unique dataset pairing 109,180 nucleus annotations with clinical biomarkers, and its comprehensive benchmark, establishing performance baselines that enable, for the first time, direct study of morphology-phenotype correlations within a single breast cancer cohort. Our benchmarks establish performance baselines for both segmentation and classification tasks, while multi-task experiments demonstrate that morphological features contain predictive signals for receptor status, offering promising directions for future investigation. By publicly releasing BC-MultiSet with comprehensive benchmarks, we provide the community with a foundational resource for developing integrated models that move beyond task-specific boundaries toward holistic histopathology analysis. Foundation models like Virchow2 excel at classification, but specialized architectures remain essential for morphology: CellViT++ leads in boundary delineation (bPQ = 56.3%) and detection

($F1_{det}$ = 80.2%), while HoVer-NeXt achieves near-comparable cell-type classification (mPQ = 40.9% vs. 41.4%) with substantially faster inference. Our multi-task learning experiments highlight the challenge of negative transfer. A staged training protocol allowed HoVer-NeXt to predict clinical biomarkers with minimal loss of segmentation performance. However, frozen morphological features showed limited discriminative power for receptor status, indicating the need for stronger supervision or task-specific architectures. External validation is limited because TCGA-BRCA provides receptor labels without nucleus annotations, whereas PanNuke and NuCLS provide nucleus annotations without receptor labels. BC-MultiSet uniquely combines both. The three-class ontology captures clinically relevant TME components. MIL approaches (ABMIL, CLAM, TransMIL) are likely to overfit this small cohort. Future work should focus on larger cohorts, increased patch sampling, and task-specific architectures.

Acknowledgments The work was supported by the Russian Science Foundation grant No 25-71-10088.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **71**(3), 209–249 (2021)
2. Goldhirsch, A., Winer, E.P., Coates, A.S., Gelber, R.D., Piccart-Gebhart, M., Thürlimann, B., Wood, W.C.: 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. Annals of Oncology **24**(9), 2206–2223 (2013)
3. Badve, S., Baehner, F.L., Gray, R.P., Childs, B.H., Maddala, T., Liu, M.L., Sparano, J.A.: Estrogen-and progesterone-receptor status in ECOG 2197: comparison of immunohistochemistry by local and central laboratories and quantitative reverse transcription polymerase chain reaction by central laboratory. Journal of Clinical Oncology **26**(15), 2473–2481 (2008)
4. Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F.L., Penault-Llorca, F., Perez, E.A., Thompson, E.A., Symmans, W.F., Richardson, A.L., Brock, J., Criscitiello, C., Bailey, H., Ignatiadis, M., Floris, G., Sparano, J., Kos, Z., Nielsen, T., Rimm, D.L., Allison, K.H., Reis-Filho, J.S., Loibl, S., Sotiriou, C., Viale, G., Badve, S., Adams, S., Willard-Gallo, K., Loi, S.: The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Annals of Oncology **26**(2), 259–271 (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778 (2016)

6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708 (2017)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
8. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Severson, K.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)
9. Baumann, E., Dislich, B., Rumberger, J.L., Nagtegaal, I.D., Martinez, M.R., Zlobec, I.: HoVer-NeXt: A Fast Nuclei Segmentation and Classification Pipeline for Next Generation Histopathology. In: Proceedings of The 7th International Conference on Medical Imaging with Deep Learning **250**, 61–86 (2024)
10. Hörst, F., Rempe, M., Becker, H., Heine, L., Keyl, J., Kleesiek, J.: CellViT++: Energy-Efficient and Adaptive Cell Segmentation and Classification Using Foundation Models. Computer Methods and Programs in Biomedicine, 109206 (2026)
11. Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z.: A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging **39**(5), 1380–1391 (2020)
12. Amgad, M., Atteya, L.A., Hussein, H., Mohammed, K.H., Hafiz, E., Elsebaie, M.A.T., Alhusseiny, A.M., AlMoslemany, M.A., Elgazar, A.F., Abdulkarim, A.: NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification, localization, and segmentation in breast cancer. Gigascience **11**, giac037 (2022)
13. Gamper, J., Koohbanani, N.A., Benet, K., Khuram, A., Rajpoot, N.: PanNuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification **11435** 11–19 (2019)
14. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61–70 (2012)
15. Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G.: QuPath: Open source software for digital pathology image analysis. Sci. Rep. **7**(1), 1–7 (2017)
16. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: Proceedings of the ISBI 2009, pp. 1107–1110. IEEE (2009)
17. Li, Z., Gu, T., Li, B., Xu, W., He, X., Hui, X.: ConvNeXt-based fine-grained image classification and bilinear attention mechanism model. Applied Sciences **12**(18), 9016 (2022)
18. Wang, X., Du, Y., Yang, S., Zhang, J., Wang, M., Zhang, J., Han, X., et al.: RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. Medical Image Analysis **83**, 102645 (2023)
19. Graham, S., Vu, Q. D., Jahanifar, M., Raza, S. E. A., Minhas, F., Snead, D., Rajpoot, N.: One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. Medical Image Analysis **83**, 102685 (2023)