Tumor Cell Fraction Estimation Based on Tissue Region Segmentation and Nuclear Density

Lulu Qin[a,b], Xiao Yang[c], Zhigang Pei[d], Susan Fotheringham[b], Xianhong Xu[b,e,*], Zexuan Zhu[f,*]

[a]*College of Computer Science and Software Engineering, Shenzhen University, Nanhai Avenue 3688, Shenzhen, 518060, China*
[b]*Oxford Cancer Biomarkers Ltd, Magdalen Centre, Oxford Science Park, Oxford, OX4 4GA, UK*
[c]*GeneGenieDx Corporation, 7068 Koll Center Pkwy, Pleasanton, CA, 94566, USA*
[d]*Pathology Department, Chongqing University Jiangjing Hospital, Jiangjing District, Chongqing, China*
[e]*Department of Oncology, University of Oxford, Oxford, OX3 7DQ, UK*
[f]*School of Artificial Intelligence, Shenzhen University, Nanhai Avenue 3688, Shenzhen, 518060, China*

Abstract

Tumor cell fraction (TCF), or tumor purity, is a critical factor in cancer diagnosis, prognosis, and molecular profiling. While genomic methods provide accurate TCF estimates, they are costly, time-consuming, and lack spatial resolution. Recent whole slide image-based approaches offer a more scalable alternative but often suffer from limited interpretability and inconsistent accuracy. We propose a novel TCF estimation method based on tissue–nuclei density (TNuD), integrating tissue region segmentation and nuclear classification from hematoxylin and eosin -stained whole slide images. The method consists of a DeepLabV3+-based tissue region segmentation model and a HoVer-Net-based nuclear segmentation and classification model. The outputs of the two models are fused to construct a TNuD matrix representing the spatial density relationships between tissue and nuclei. We evaluated the proposed method against four TCF estimation baselines using expert-annotated breast and ovarian cancer datasets. The proposed TNuD-based method achieved the lowest mean squared error (MSE = 0.0214) and highest correlation with pathologist annotations (Pearson = 0.8683; Spearman = 0.8737) in breast cancer datasets. It also demonstrated promising transferability to ovarian cancer tissues. Comparative analysis also showed superior precision and interpretability over region- or nucleus-only models. The TNuD-based method effectively captures tumor heterogeneity by combining macro- and micro-level histological features. It offers a scalable, interpretable, and accurate solution for TCF estimation in digital pathology, supporting broader clinical and translational oncology applications.

Keywords: Tumor cell fraction, Digital pathology, Whole slide image, Tissue segmentation, Nuclear segmentation and classification

Introduction

Cancer is a complex and heterogeneous group of diseases characterized by uncontrolled cell proliferation, invasion of surrounding tissues, and the potential for metastasis to distant organs Siegel et al. (2023); Shi et al. (2024). Tumor microenvironment (TME) is a complex ecosystem composed of tumor cells and their surrounding cellular, molecular, vascular, and extracellular matrix components, as illustrated in Figure 1 Anderson and Simon (2020); Crosby et al. (2022). Interactions among these components drive tumor growth, metastasis, and response to therapy, making the TME a key target for the development of new therapeutic strategies and improvement of treatment outcomes Bejarano et al. (2021); Zhang et al. (2024).

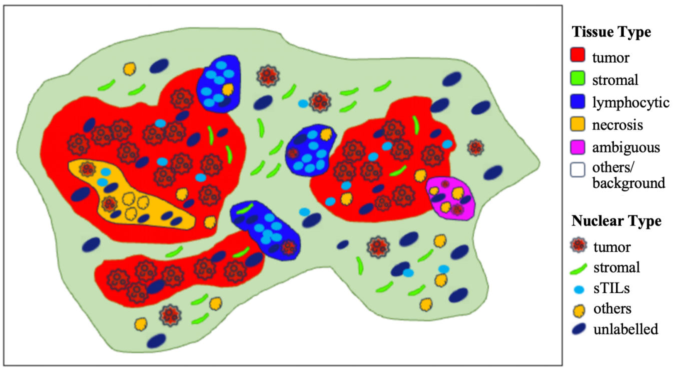


Figure 1: A schematic diagram illustrating the distribution of different tissue regions and corresponding nuclear types in a WSI. Tissue region types (e.g., tumor, stromal, lymphocytic, necrosis, ambiguous regions) and nuclear types (e.g., tumor nuclei, stromal nuclei, sTILs, others, unlabelled nuclei) are represented by distinct colors, as indicated in the legend.

Tumor cell fraction (TCF), or tumor purity, refers to the proportion of cancer cells within a tumor sample Brendel et al. (2022). TCF is an important measurement for understanding TME, as it provides insights into TME composition and is a key factor influencing tumor biology, immune evasion, therapeutic response, and clinical outcomes. In tissue sample DNA sequencing, a high TCF provides a more accurate representa-

*Corresponding authors.
E-mail addresses: 2050271008@email.szu.edu.cn (L. Qin), xyang@genegeniedx.com (X. Yang), 17784255157@cqu.edu.cn (Z. Pei), Susan.Fotheringham@oxfordbio.com (S. Fotheringham), xianhong.xu@oncology.ox.ac.uk (X. Xu), zhuzx@szu.edu.cn (Z. Zhu).

Preprint submitted to Expert Systems with Applications *May 23, 2025*

tion of the tumor's genomic characteristics, while a low TCF can dilute biomarker signals, increasing the risk of false negatives and complicating treatment decisions Cheng et al. (2020); Koo and Rhee (2021); Yu et al. (2023). TCF also serves as a prognostic indicator, e.g., lower TCF in colorectal cancer is associated with poorer survival rates West et al. (2010), while higher TCF in stage I lung adenocarcinoma correlates with better post-surgery outcomes Jeon et al. (2022).

TCF evaluation typically involves computational analysis of integrated genomic data (epigenomic, genomic, and transcriptomic) or examination of digital pathology whole slide images (WSIs) of the tissue samples Haider et al. (2020). Genomic sequencing methods, including transcriptomic Revkov et al. (2023) as well as somatic alteration and methylation-based approaches Zheng et al. (2014); Locallo et al. (2019), provide comprehensive molecular insights and thus generally offer more precise TCF estimates Tbeileh et al. (2023). However, the transcriptomic approaches struggle with stromal or immune infiltration and variability in RNA sequencing protocols Benelli et al. (2018); Choi et al. (2023). The somatic alteration and methylation-based approaches can be constrained by sample type, sequencing depth, and the presence of matched normal references Zheng et al. (2014); Wei et al. (2021). In addition, the high costs and lengthy data processing times of these approaches limit their applicability in rapid clinical testing scenarios.

In contrast, tissue image-based approaches leverage histopathological structures to estimate TCF, representing a cost-effective alternative and emerging as a preferred option Augustine (2022). However, they are limited by staining variability, the need for extensive annotations, and segmentation difficulties in immune-infiltrated or densely packed regions Schoenpflug et al. (2025); Osinski et al. (2022); Brendel et al. (2022); Choi et al. (2023). Both types of approaches have advanced TCF estimation, yet challenges remain regarding cross-institutional variability, interpretability, and data availability Haider et al. (2020); Menzel et al. (2023). Visual analysis of tissue images by trained pathologists remains widely used, but faces challenges such as subjectivity, variability in image quality, difficulty in distinguishing tumor cells, and inefficiencies in high-throughput image analysis Haider et al. (2020). More recently, artificial intelligence (AI)-driven computational approaches for WSI-based TCF estimation, including tissue region-based Brendel et al. (2022); Su et al. (2022); Cheng et al. (2025) and nuclear detection approaches Liu et al. (2020); Sun et al. (2021); Priego-Torres et al. (2022); Sakamoto et al. (2022); Choi et al. (2023); Liu et al. (2024a); Kang et al. (2024), have been developed to address these limitations. Many of these methods rely on black-box AI algorithms, whose lack of interpretability hinders clinical adoption. In contrast, nuclear detection methods enable TCF evaluation at the nuclear level, providing greater detail and interpretability Silva et al. (2022). However, challenges remain in achieving accurate nuclear detection and classification, especially given the complexity and scale of WSI datasets Gerardin et al. (2024). These limitations highlight the need for advanced models that integrate scalability, interpretability, and high accuracy to address the challenges of TCF estimation effectively.

Table 1: Summary of Abbreviations

	Description
TME	Tumor Microenvironment
TCF	Tumor Cell Fraction
H&E	Hematoxylin-Eosin
WSI	Whole Slide Image
AI	Artificial Intelligence
CNN	Convolutional neural network
TNuD	Tissue-Nuclei Density
TRS	Tissue Region Segmentation
NuSC	Nuclei Segmentation and Classification
ROI	Regions of Interest
TC	Tile Classification
NuCLS	Nucleus Classification, Localization, and Segmentation
BCSS	Breast Cancer Semantic Segmentation
NP	Nuclear Pixel
HoVer	Horizontal and Vertical distance maps
NC	Nuclear Classification
MSE	Mean Squared Error
TILs	Tumor-Infiltrating Lymphocytes
GT	Ground Truth
MIL	multiple instance learning

In this article we propose a Tissue-Nuclei Density (TNuD)-based TCF estimation method, which offers a comprehensive and clinically interpretable approach by integrating both tissue- and nuclei-level information from the WSI. Central to this method is the TNuD model, which quantifies the density relationships between tissue region types and nuclei by calculating nuclear areas and counts within specific tissue regions. Specifically, tissue- and nuclei-level information is obtained through Tissue Region Segmentation (TRS) and Nuclei Segmentation and Classification (NuSC). To evaluate the proposed method, we conducted an investigational clinical study to collected 54 breast and 52 ovarian cancer tissue samples and construct Hematoxylin-Eosin (H&E) WSI datasets with detailed tumor region of interest (ROI) annotations and TCF scores manually estimated by experienced pathologists. The abbreviations used throughout this paper are summarized in Table 1 for clarity and consistency.

Methods

The proposed TNuD-based TCF estimation approach combines TRS-based and NuSC-based methods. The framework of this approach is illustrated in Figure 2. First, an H&E WSI image undergoes preprocessing, which includes color normalization to mitigate staining variations and the extraction of a tissue region mask to exclude non-tissue areas. Next, a sliding window approach with overlapping regions divides the WSI into smaller image patches, enabling localized analysis while preserving spatial context. In the subsequent steps, a trained TRS model predicts the tissue region type for each pixel within the patches. These predictions are then aggregated to generate a

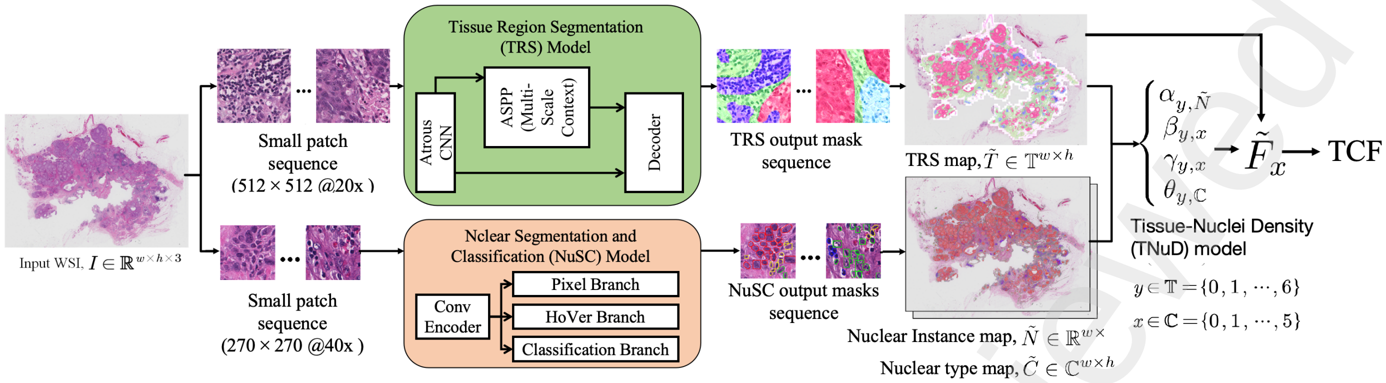

Input WSI, $I \in \mathbb{R}^{w\times h\times 3}$
Small patch sequence (512 × 512 @20x)
Small patch sequence (270 × 270 @40x)
Tissue Region Segmentation (TRS) Model
Atrous CNN
ASPP (Multi-Scale Context)
Decoder
Nclear Segmentation and Classification (NuSC) Model
Conv Encoder
Pixel Branch
HoVer Branch
Classification Branch
TRS output mask sequence
NuSC output masks sequence
TRS map, $\tilde{T} \in \mathbb{T}^{w\times h}$
Nuclear Instance map, $\tilde{N} \in \mathbb{R}^{w\times}$
Nuclear type map, $\tilde{C} \in \mathbb{C}^{w\times h}$
$\alpha_{y,\tilde{N}}$ $\beta_{y,x}$ $\gamma_{y,x}$ $\theta_{y,\mathbb{C}}$
$\tilde{F}_x$
TCF
Tissue-Nuclei Density (TNuD) model
$y \in \mathbb{T} = \{0, 1, \cdots, 6\}$
$x \in \mathbb{C} = \{0, 1, \cdots, 5\}$

Figure 2: The workflow of obtaining the TNuDM.

comprehensive tissue region map for the entire WSI. Simultaneously, a trained NuSC model detects, segments, and classifies individual nuclei within each patch. The outputs from all patches are combined to produce a nuclear instance mask and a nuclear type mask for the WSI, ensuring a precise representation of nuclear distributions. Finally, the tissue region maps ($\tilde{T}$) and nuclear maps ($\tilde{N}$ and $\tilde{C}$) are fed into a TNuD model, which quantifies nuclear density relationships across different tissue regions using a set of relationship matrices.

Building on the authors' previous study Wang et al. (2023), a DeepLabV3+ model was implemented in the TRS model Chen et al. (2018); Vykopal et al. (2023); Liu et al. (2024b). This architecture leverages the capability of Convolutional neural network (CNN) and Atrous Spatial Pyramid Pooling (ASPP) to capture global contextual information, thereby enhancing segmentation efficacy. HoVer-Net Graham et al. (2019) was adopted in the NuSC model. It is a CNN-based framework specifically designed for the concurrent segmentation and classification of nuclear instances in histopathological images. It was demonstrated that incorporating tissue region information as additional parameters significantly improved nuclei classification accuracy Wang et al. (2023); Hörst et al. (2024). The TNuD model was designed based on our observations of the TME to integrate intricate nuclear density relationships within the TME into TCF analysis, resulting in a more precise, comprehensive, and clinically interpretable TCF estimation. This represents the core innovation of the proposed method, and its design is described here.

TNuD Model Design

TME exhibits a distinct correlation between tissue type and nuclear density, as illustrated in Figure 1. Proliferative regions, such as tumors and lymphocytic infiltrates, exhibit high nuclear density, whereas non-proliferative or necrotic areas show markedly lower nuclear density. Furthermore, specific tissue regions are predominantly composed of certain nucleus types, reflecting their functional or pathological characteristics. However, these regions usually contain a mixture of other nucleus types, reflecting the inherent complexity and heterogeneity of the TME. The TNuD model was proposed to capture and quantify the density relationships between various tissue types and their corresponding nuclei. Its mathematical definition is provided as follows. For clarity, Table 2 summarizes all parameter symbols and their definitions.

Table 2: Summary of Mathematical Symbols

Symbol	Description
$I \in \mathbb{R}^{w\times h\times 3}$	WSI with width w, height h, and RGB channels
$\mathbb{T}$	Set of all tissue types.
$\mathbb{C}$	Set of all nuclear types.
$T \in \mathbb{T}^{w\times h}$	Tissue region map assigning a tissue type to each pixel
$N \in \mathbb{R}^{w\times h}$	Nuclear instance map assigning a unique label to each nucleus
$C \in \mathbb{C}^{w\times h}$	Nuclear type map indicating the nuclear type for each pixel
x	one of nuclear types
y	one of tissue types
C_x	Binary mask for nuclear type x
T_y	Binary mask for tissue type y
N_x	Nuclear instance mask for a specific nuclear type x
$\odot$	Element-wise product operation
F_x	Nuclei purity fraction for nuclear type x
$\lvert \cdot \rvert_+$	Cardinality of a matrix, counting unique nonzero elements
$\tilde{T}, \tilde{N}, \tilde{C}$	Estimated tissue region, nuclear instance, and nuclear type maps.
$\tilde{F}_x$	Estimated nuclei purity fraction for nuclear type x
$\lVert \cdot \rVert_0$	Zero-norm of a matrix (count of nonzero elements)
$\alpha_{y,\tilde{N}}$	Proportion of tissue area occupied by nuclear pixels
$\beta_{y,x}$	Proportion of nuclear pixels within tissue area $\tilde{T}_y$ that are of type x
$\gamma_{y,x}$	Density of nuclei of type x in tissue region $\tilde{T}_y$
$\theta_{y,\mathbb{C}}$	Average number of nuclei per unit nuclear area in tissue region $\tilde{T}_y$
$\bar{\alpha}_{y,N}$	Averaged $\alpha_{y,\tilde{N}}$ over multiple WSIs
$\bar{\beta}_{y,x}$	Averaged $\beta_{y,x}$ over multiple WSIs
$\bar{\gamma}_{y,x}$	Averaged $\gamma_{y,x}$ over multiple WSIs
$\bar{\theta}_{y,\mathbb{C}}$	Averaged $\theta_{y,\mathbb{C}}$ over multiple WSIs
m	Total number of WSIs
$\tilde{T}_y$	Estimated tissue type mask of type y

Let a WSI be denoted as $I \in \mathbb{R}^{w\times h\times 3}$, where w and h represent the image's width and height, respectively, and the three channels correspond to the RGB color space. A WSI can be featured by three maps, i.e., the tissue region map T, the nuclear instance map N, and the nuclear type map C. Specifically, the tissue region map $T \in \mathbb{T}^{w\times h}$ assigns a tissue type to each pixel. If $T_{i,j} = y$, it indicates that the pixel at location (i, j) belongs to tissue type y, with $y = 0$ denoting the absence of a specific tissue type. The nuclear instance map $N \in \mathbb{R}^{w\times h}$ assigns a unique label to each nucleus, where $N_{i,j} = k$ signifies that the pixel at (i, j) corresponds to the k-th nucleus, with k as a positive integer serving as its unique identifier. If a pixel does not belong to any nucleus, k is set to 0. Lastly, the nuclear type map $C \in \mathbb{C}^{w\times h}$ identifies the nuclear type for each pixel. If $C_{i,j} = x$, the pixel at location (i, j) is assigned nuclear type x, with $x = 0$ indicating that the pixel does not correspond to any nucleus.

For conciseness and clarity, we introduce the matrix C_x to identify the pixels belonging to nuclear type x as:

$$(C_x)_{i,j} = \begin{cases} 1, & \text{if } C_{i,j} = x \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Similarly, the tissue type indicator matrix T_y identifies the pixels corresponding to tissue type y, i.e.,

$$(T_y)_{i,j} = \begin{cases} 1, & \text{if } T_{i,j} = y \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Furthermore, the nuclear instance mask for a specific type x, represented by N_x, refines the nuclear instance mask N by including only the nuclei of type x, i.e., $N_x = N \odot C_x$, where $\odot$ is an element-wise product.

The nuclei purity fraction is defined as the ratio of a specific nuclear type to the total nuclear count within a WSI. Particularly, the nuclei purity fraction of nuclear type x in a WSI can be calculated as follows:

$$F_x = \frac{|N_x|_+}{|N|_+} = \frac{|\{k \mid k \in N_x, k \neq 0\}|}{|\{k \mid k \in N, k \neq 0\}|} \tag{3}$$

where $|\cdot|_+$ indicates the cardinality of a given matrix, i.e., the unique elements of the matrix, excluding 0. Thus, $|N|_+$ represents the total number of nuclei in the WSI, and $|N_x|_+$ denotes the number of nuclei of type x.

Based on the above definitions, the calculation of F_x is straightforward given known T, N and C. However, in practice, it is challenging to obtain accurate values of T, N and C, due to the large-scale nature of WSIs, heterogeneous staining, morphological variability, overlapping nuclei, and the subjective nature of manual annotation. To address this, we can use TRS model to obtain the estimated T, i.e., $\tilde{T}$, and NuCS model to attain the estimated N and C, i.e., $\tilde{N}$ and $\tilde{C}$, as shown in Figure 2. Note that estimating N and C is much more challenging than estimating T, because nuclei are smaller and harder to distinguish. Moreover, NuCS is much more time-consuming than TRS. To further mitigate the computational burden, we train a TNuD model to capture the density relationship between the tissue region matrix $\tilde{T}$ and the nuclear matrices $\tilde{N}$ and $\tilde{C}$ with training data, enabling the estimation of N and N_x using only tissue region information from $\tilde{T}$ in the inference process. Given a WSI, we can estimate F_x as follows:

$$\tilde{F}_x = \frac{|\tilde{N}_x|_+}{|\tilde{N}|_+} = \frac{\sum_y \left(\|\tilde{T}_y\|_0 \cdot \alpha_{y,\tilde{N}} \cdot \beta_{y,x} \cdot \gamma_{y,x}\right)}{\sum_y \left(\|\tilde{T}_y\|_0 \cdot \alpha_{y,\tilde{N}} \cdot \theta_{y,\mathbb{C}}\right)} \tag{4}$$

where $\|\cdot\|_0$ calculates the zero-norm of a matrix, i.e., the number of non-zero elements in the given matrix. The parameters $\alpha_{y,\tilde{N}}$, $\beta_{y,x}$, $\gamma_{y,x}$, and $\theta_{y,\mathbb{C}}$ are specific to the TNuD model, characterizing the spatial distribution and density of nuclei within tissue regions. They are fitted with $\tilde{T}_y$, $\tilde{N}$, $\tilde{N}_x$, and $\tilde{C}_x$ obtained in the training data. After training, the parameters are fixed and used in the inference stage to avoid involving the time-consuming estimation of N and N_x. Particularly, these parameters are defined as follows:

- $\alpha_{y,\tilde{N}}$ quantifies the proportion of tissue area $\tilde{T}_y$ occupied by nuclear pixels, providing a measure of overall nuclear density within a specific tissue region and indicating how much of the tissue is composed of nuclei:

$$\alpha_{y,\tilde{N}} = \frac{\|\tilde{N} \odot \tilde{T}_y\|_0}{\|\tilde{T}_y\|_0} \tag{5}$$

- $\beta_{y,x}$ represents the proportion of nuclei pixels within tissue area $\tilde{T}_y$ that are of type x. It indicates the relative abundance of a specific nucleus type among all nuclei present in the tissue region, providing insight into the distribution of different nuclear types within the tissue.

$$\beta_{y,x} = \frac{\|\tilde{C}_x \odot \tilde{T}_y\|_0}{\|\tilde{N} \odot \tilde{T}_y\|_0} \tag{6}$$

- $\gamma_{y,x}$ denotes the nuclei density of type x within the tissue area $\tilde{T}_y$. By taking the reciprocal of the average area per nucleus, i.e., $f(\tilde{C}_x)/p(\tilde{N}_x)$, it provides the number of nuclei per unit area for that specific nucleus type. This metric indicates how densely the nuclei of type x are packed within the tissue region. A higher value of $\gamma_{y,x}$ implies a greater concentration of nuclei of that type.

$$\gamma_{y,x} = \frac{|\tilde{N}_x \odot \tilde{T}_y|_+}{\|\tilde{C}_x \odot \tilde{T}_y\|_0} \tag{7}$$

- $\theta_{y,\mathbb{C}}$ indicates the average number of nuclei per unit nuclear area within a tissue region. It reflects the nuclear density within the nuclear regions of the tissue, offering a focused perspective on cellular distribution and potential changes in tissue morphology.

$$\theta_{y,\mathbb{C}} = \frac{\sum_x \left(\|\tilde{N}_x \odot \tilde{T}_y\|_0 \cdot \gamma_{y,x}\right)}{\|\tilde{N} \odot \tilde{T}_y\|_0} \tag{8}$$

According to Eq. (4) and the process outlined in Figure 2, a specific TNuD model can be built for a given WSI. However, to accurately predict cell fractions in previously unseen

WSIs during future inference, it is essential to establish a general TNuD model to capture typical patterns of nuclear density, cell type distribution, and tissue architecture across diverse samples, thereby ensuring broader applicability to a wide range of WSIs. We can generalize the TNuD model by averaging the model parameters associated with a series of diverse WSIs:

$$\bar{\alpha}_{y,N} = \frac{1}{m}\sum_{i=1}^{m} \alpha^{i}_{y,\tilde{N}}, \tag{9}$$

$$\bar{\beta}_{y,x} = \frac{1}{m}\sum_{i=1}^{m} \beta^{i}_{y,x}, \tag{10}$$

$$\bar{\gamma}_{y,x} = \frac{1}{m}\sum_{i=1}^{m} \gamma^{i}_{y,x}, \tag{11}$$

$$\bar{\theta}_{y,\mathbb{C}} = \frac{1}{m}\sum_{i=1}^{m} \theta^{i}_{y,\mathbb{C}} \tag{12}$$

where $\alpha^{i}_{y,\tilde{N}}$, $\beta^{i}_{y,x}$, $\gamma^{i}_{y,x}$, and $\theta^{i}_{y,\mathbb{C}}$ denote the corresponding parameters obtained in the i-th WSI, and m is the total number of WSIs. The final $\tilde{F}_x$ for the nuclei purity fraction estimation of nuclei type x is then rewritten as:

$$\tilde{F}_x = \frac{\sum_{y} \left(\|\tilde{T}_y\|_0 \cdot \bar{\alpha}_{y,\tilde{N}} \cdot \bar{\beta}_{y,x} \cdot \bar{\gamma}_{y,x} \right)}{\sum_{y} \left(\|\tilde{T}_y\|_0 \cdot \bar{\alpha}_{y,\tilde{N}} \cdot \bar{\theta}_{y,\mathbb{C}} \right)} \tag{13}$$

As shown in Eq. (13), to calculate the TCF of a cancer type x in a new WSI, it is only necessary to obtain $\tilde{T}_y$ via TRS as the other parameters have already been fitted with training data.

Data Collection and Model Training

The proposed TNuD-based TCF estimation method was evaluated on real-world WSI datasets shown in Table 3 and compared with other state-of-the-art TCF estimation techniques. Derived mainly from breast and ovarian cancer specimens and annotated for tissue regions, nuclear details, and TCF estimates, these datasets were curated to support specific computational pathology tasks such as TRS, NuSC, ROI segmentation, and TCF estimation.

The BCSS (Breast Cancer Semantic Segmentation, https://bcsegmentation.grand-challenge.org/) dataset, sourced from H&E-stained breast cancer images across 18 institutions from The Cancer Genome Atlas (TCGA), was used to train the TRS model. We streamlined this dataset to six super tissue region classes (“Tumor,” “Stroma,” “Lymphocyte,” “Necrosis,” “Ambiguous,” “Others”) as per preprocessing protocols detailed in Amgad et al. Amgad et al. (2019). The dataset was divided into an 8:2 training-test split. Patches were extracted at 512 × 512 resolution, halved to 256 pixels during cropping. Augmentation techniques including mean subtraction, random resizing, and horizontal flipping were used to enhance training diversity. For inference, a sliding window method ensured uniform patch resolution, minimizing boundary errors and improving segmentation accuracy.

The NuCLS (Nucleus Classification, Localization, and Segmentation, https://sites.google.com/view/nucls) dataset features over 220,000 annotations including nuclear instance masks and type labels. A pathologist-verified subset (“corrected single rater dataset”) was utilized to train the NuSC model. Original cell class annotations were grouped into superclasses -“tumor”, “stromal”, “sTILs”, “others”, “unlabeled.” - to ensure clinical relevance and balanced training. The dataset was split in a 7:2:1 ratio for training, validation, and test.

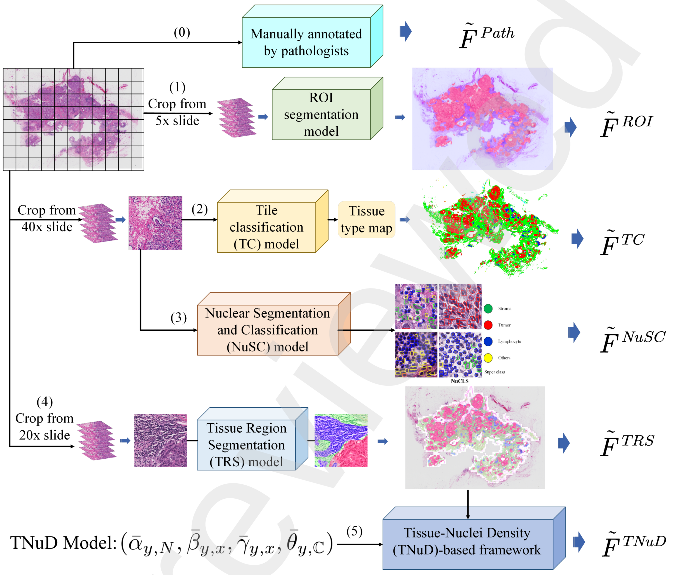

(0)
Manually annotated by pathologists
$\tilde{F}^{Path}$
(1)
Crop from 5x slide
ROI segmentation model
$\tilde{F}^{ROI}$
Crop from 40x slide
(2)
Tile classification (TC) model
Tissue type map
$\tilde{F}^{TC}$
(3)
Nuclear Segmentation and Classification (NuSC) model
Stroma
Tumor
Lymphocyte
Others
Super class
NuCLS
$\tilde{F}^{NuSC}$
(4)
Crop from 20x slide
Tissue Region Segmentation (TRS) model
$\tilde{F}^{TRS}$
TNuD Model: $(\bar{\alpha}_{y,N}, \bar{\beta}_{y,x}, \bar{\gamma}_{y,x}, \bar{\theta}_{y,\mathbb{C}})$
(5)
Tissue-Nuclei Density (TNuD)-based framework
$\tilde{F}^{TNuD}$

Figure 3: Overview of six different methods to estimate the nuclei proportion score.

We constructed the Cancer-ROI, Breast-WSIs, and Ovarian-WSIs datasets from clinical samples of breast and ovarian cancer resections to evaluate the TNuD-based TCF estimation. Slides were scanned into WSIs at 40× magnification, then annotated by experienced pathologists to produce tissue region masks and TCF scores following structured protocols.

The Cancer-ROI dataset includes 10 breast cancer WSIs with tumor regions of interest (ROI) annotations. We extracted 512 × 512 image patches from 5× magnification images using a sliding window method with 128-pixel overlap, and divided them into training, validation, and test sets at a 7:2:1 ratio.

The Breast-WSIs dataset, consisting of 54 breast WSIs, serves as the reference dataset for evaluating both ROI detection and TCF estimation. Its combination with the BCSS and NuCLS datasets was the base for the development of breast cancer TNuD model. The Ovarian-WSIs dataset, consisting of 52 ovarian WSIs was used for evaluating the TCF estimation and testing the transferability of the TNuD-based method across cancer types.

Performance Evaluation Experiment

As shown in Figure 3, we evaluated the performance of the proposed TNuD-based TCF estimation method against the pathologist’s manual estimation method and four other TCF estimation methods:

- $\tilde{F}^{Path}$: Represents the TCF estimated by pathologists through visual assessment. This benchmark reference, despite being the clinical standard, is subject to variability

Table 3: Overview of the datasets used in this study.

Dataset	Task	#Img	Image Size	Annotations	#Instances	#Types
BCSS	TRS	151	1222×1036 to 7360×6813	Multi-type masks	20,340	21
NuCLS	NuSC	3,944	287×292 to 830×765	Multi-type, instance masks	222,396	12
Cancer-ROI	Segmentation	20	WSIs	Tumor region masks	\	1
Breast-WSIs	Prediction	54	WSIs	WSI-based TCF scores	54	\
Ovarian-WSIs	Prediction	52	WSIs	WSI-based TCF scores	52	\

The notation #Img signifies the total number of images or WSIs included in the dataset. #Instances records the total number of tissue region or nuclei instances contained within the dataset. #Types refers to the number of distinct annotated tissue region or nuclear types represented.

due to factors such as staining inconsistencies and observer fatigue.

- $\tilde{F}^{ROI}$: Represents the TCF estimation calculated as the ratio of the tumor region area to the total tissue area in the WSI using a ROI-based method, similar to those used by Oner et al. Oner et al. (2022) The key difference is that we use a lower magnification (5×) image to capture large-scale tumor distribution.

- $\tilde{F}^{TC}$: Represents the TCF estimation calculated as the ratio of tumor patches to all tissue patches within the WSI at 40× magnification using a tile classification (TC)-based method, similar to those used by Brendel et al. Brendel et al. (2022) and Fu et al. Fu et al.. The key difference is that we adopted a supervised approach for training the tile image classification model, using EfficientNet Tan and Le (2020).

- $\tilde{F}^{NuSC}$: Represents the TCF estimation calculated as the ratio of tumor nuclei to total nuclei in the tissue sample WSI at 40× magnification, with nuclei detected by the NuSC model within our TNuD-based framework.

- $\tilde{F}^{TRS}$: Represents the TCF estimation calculated as the ratio of the tumor area to the total tissue area in the sample tissue WSI at 20× magnification, with tissue and tumor region mask output generated by the TRS model within our TNuD-based framework.

The objective of the performance evaluation was to assess the correlation between the predicted TCF values and the ground truth annotations using several evaluation metrics, including mean squared error (MSE), Pearson and Spearman correlation coefficients, and linear regression analysis. It was carried out on the breast and ovarian datasets respectively. It is important to note that the key hyperparameter of the TNuD-based framework was optimized on breast cancer data. The experiment results with ovarian dataset served as an exploration of the model's transferability across different cancer types.

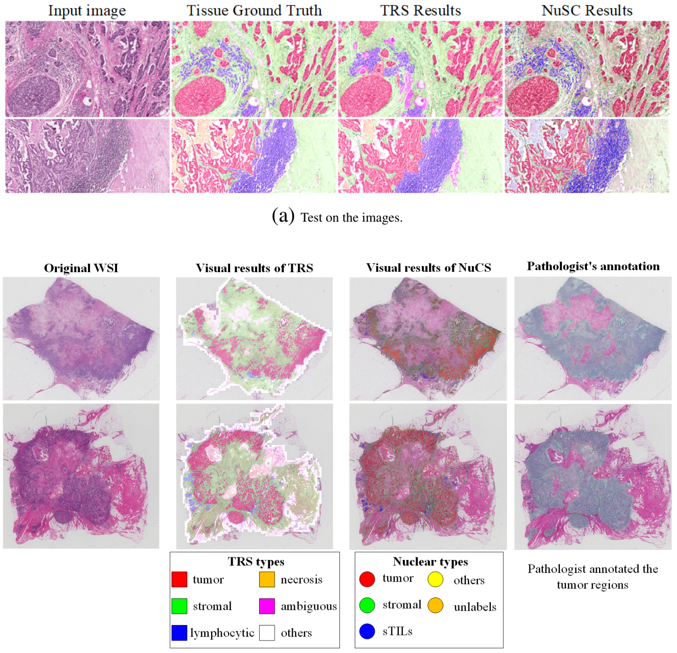
Results

Breast Cancer Results

The performance of the trained TRS and NuSC models were first examined. Quantitative results for both models are summarized in Table 4. The TRS model achieves high accuracy across tissue types, while the NuSC model performs well in nuclear segmentation and yields moderate classification results (e.g., 0.6707 Dice for tumor nuclei). Figure 4 shows two examples of the outputs of the TRS and NuSC models at the patch image-level and WSI-level, respectively. Figure 4 (a) shows that the TRS model closely approximates the ground truth (GT). A comparative analysis of the NuSC and TRS results reveals a strong correlation between tissue regions and their corresponding nuclear types, validating the anticipated consistency in their spatial distribution. These tissue regions frequently accommodate a varied assortment of nuclear types and that nuclear density within a region typically arises from a blend of multiple nuclear types rather than a single type. Furthermore, areas abundant in lymphocyte nuclei often neighbor or intersect tumor regions, potentially causing misclassification of tumor nuclei as lymphocytes, particularly when nuclear morphology and density alone are inadequate for precise differentiation.

Observing the WSI-level examples in Figure 4 (b), a high density of tumor nuclei within the tumor and lymphocytic regions, showing strong concordance with the TRS results and pathologist annotations. Some slight discrepancies in nuclei boundary delineations observed were linked to the inherent complexity and heterogeneity of the tissue structures in the images, especially in ambiguous areas where tissue types overlap. These discrepancies may stem from unclear manual annotations or the models' limitations in differentiating subtle morphological features of the cells.

The TCF estimation performance evaluation results are presented in Table 5 and Figure 5. The TNuD-based TCF workflow consistently outperforms all other methods across multiple metrics. It achieves the lowest MSE of 0.0214, reflecting minimal prediction error, and demonstrates the highest Pearson (0.8683) and Spearman (0.8737) correlations, indicating strong alignment with pathologist annotations. Linear regression analysis



(a) Test on the images.

(b) Test on the WSIs.

Figure 4: Examples of comparison of the outputs of TRS and NuSC models with pathologist annotations.

Table 4: Performance of TRS and NuSC Models Across Tissue and Nuclear Classes.

	TRS		NuSC NC	NuSC NP	
	Acc	Dice	Dice	Acc	Dice
Tumor	0.9111	0.8406	0.6707	N/A	N/A
Stroma	0.8490	0.8181	0.3882	N/A	N/A
Lymphocytic	0.9279	0.7413	0.6313	N/A	N/A
Necrosis	0.9780	0.7781	N/A	N/A	N/A
Others	0.9400	0.6723	0.1440	N/A	N/A
Background	N/A	N/A	0.8846	N/A	N/A
Nuclei	N/A	N/A	N/A	0.8527	0.7861

shows a slope of 0.9332 and an intercept of 0.0808, suggesting minimal bias, while the R-value of 0.7540 and low standard error (0.0769) underscore the method's precision and explanatory power. These findings are further supported by the scatter plot in Figure 5, where TNuD-based predictions closely cluster around the line of perfect correlation.

The ROI-based approach, while yielding competitive results with an MSE of 0.0232, Pearson (0.8315), and Spearman (0.8246) correlations, does not capture fine-grained nuclear details. Although it offers the advantage of rapid estimation at low magnification, this limitation reduces its overall precision compared to the TNuD-based method. The NuSC-based and TC-based methods exhibit higher MSE values and lower correlation coefficients, indicating greater variability and less reliable TCF predictions. The NuSC-based method, for instance, has a correlation coefficient of 0.7549, and its higher MSE suggests frequent inaccuracies. The TC-based method, with a Pearson correlation of 0.8113, shows a pronounced overestimation bias, as reflected in its slope of 1.2440. Both methods, as shown in Figure 5 (f), demonstrate a tendency to overestimate TCF, leading to less consistent predictions.

The TRS-based method, although demonstrating reasonable correlation with pathologist annotations, exhibits greater variability in its predictions, as indicated by its higher MSE and the broader spread observed in the scatter plot (Figure 5 (e)). While it captures general trends in TCF, the TRS-based method is less reliable in complex tissue regions, where deviations from pathologist estimates become more pronounced, particularly in higher TCF values. In comparison, the TNuD-based method improves upon the TRS-based approach by incorporating tissue region segmentation and tumor nuclear density mapping. This multi-scale integration enhances the accuracy of predictions, better captures the heterogeneity of tumor tissues, and addresses the cellular details missing in the TRS method.

In conclusion, the testing results demonstrate that the TNuD-based approach delivered highly accurate and reliable TCF predictions, outperforming other methods in error rate, correlation with pathologist annotations, and predictive power.

Ovarian Cancer Results

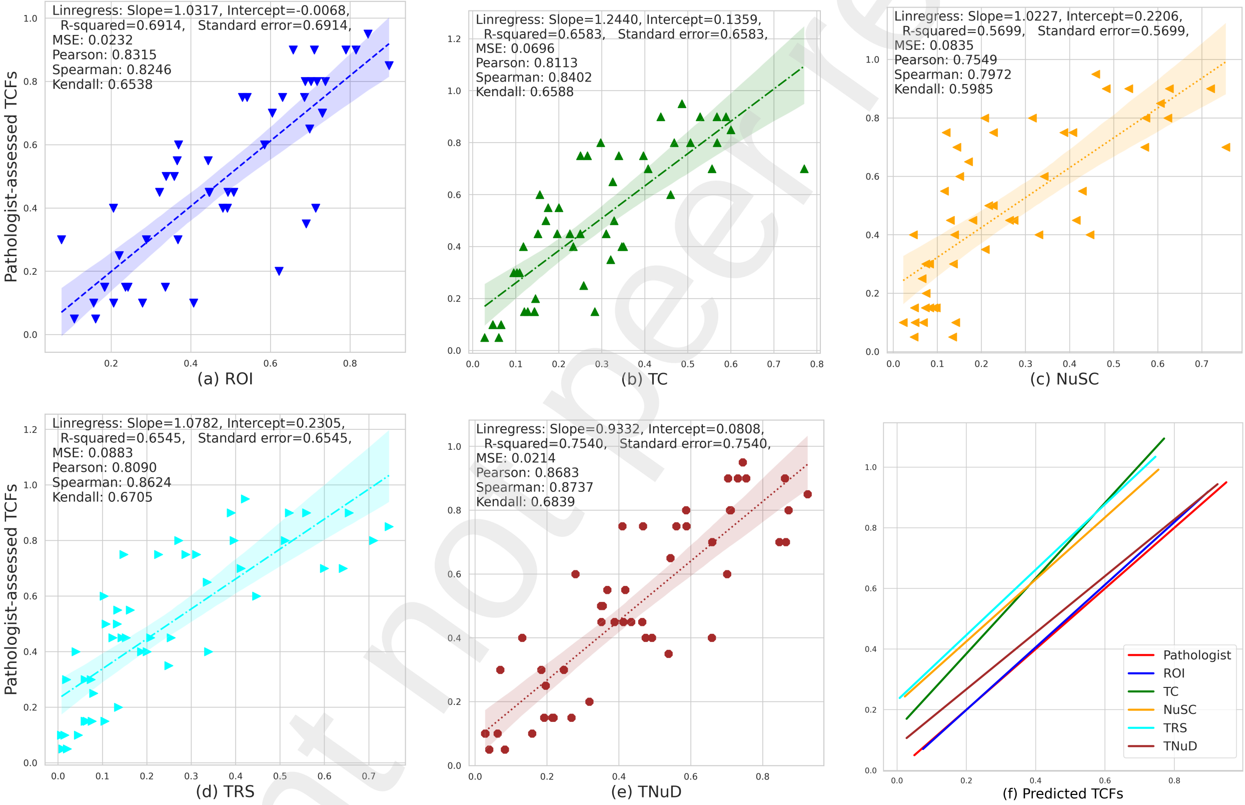
The TCF estimation performance evaluation results for the ovarian dataset are shown in Table 6 and Figure 6. Interestingly, the TC-based method demonstrates slightly stronger correlation metrics (Pearson: 0.7630, Spearman: 0.7703) than the TNuD-based workflow, alongside a very similar MSE (0.0483). However, the linear regression slope (1.1661) indicates a clear overestimation bias, and the relatively low R-value (0.5821) coupled with a higher standard error (Std Err: 0.1225) suggests notable variability in predictions. These factors indicate that while the TC-based model aligns with general trends, it does so with less consistency, limiting its practical reliability.

The TNuD-based workflow demonstrates robust performance on ovarian cancer, with an MSE of 0.0481. Its Pearson (0.7599) and Spearman (0.7670) correlations are relatively strong, though they remain somewhat lower than the high correlations observed previously in breast cancer. Nevertheless, the TNuD-based method shows important advantages in terms of consistency and minimal bias, as evidenced by the linear regression slope (0.9973), intercept (0.0787), higher R-value (0.5775), and lower standard error (0.1120). These findings suggest that the TNuD-based model, despite slightly lower correlation metrics, provides more stable and less biased predictions than the TC-based method. Given that the TNuD model was trained exclusively on breast cancer data, further optimization with an ovarian cancer-specific dataset will be required to fully realize the model's performance potential.

The ROI-based method demonstrates moderate performance in ovarian cancer, with an MSE of 0.0482. Its Pearson (0.7257) and Spearman (0.7207) correlations, along with a regression slope of 0.9201, indicate that the method can roughly capture general tumor distribution at low magnification (5×). However, because it relies on macroscopic tumor segmentation, this approach may not fully capture fine-grained tumor heterogene-

Table 5: Comparison of the correlation between the predicted TCF of different estimation methods and pathologists' annotations on breast cancer.

Model	MSE↓	Pearson↑	Spearma↑	Linear			
				Slope	Intercept	R-values ↑	Std Err ↓
ROI	0.0232	0.8315	0.8246	1.0317	-0.0068	0.6914	0.0995
TC	0.0696	0.8113	0.8402	1.2440	0.1359	0.6583	0.1294
NuSC	0.0835	0.7549	0.7972	1.0227	0.2206	0.5699	0.1282
TRS	0.0883	0.8090	0.8624	1.0782	0.2305	0.6545	0.1131
TNuD	**0.0214**	**0.8683**	**0.8737**	0.9332	0.0808	**0.7540**	**0.0769**



(a) ROI (b) TC (c) NuSC

(d) TRS (e) TNuD (f) Predicted TCFs

Figure 5: The visualization of the linear regression model between TCF predicted by different methods and pathologists' annotations on breast cancer.

ity, leading to lower accuracy compared to the TC-based and TNuD-based approaches, which use higher-resolution images.

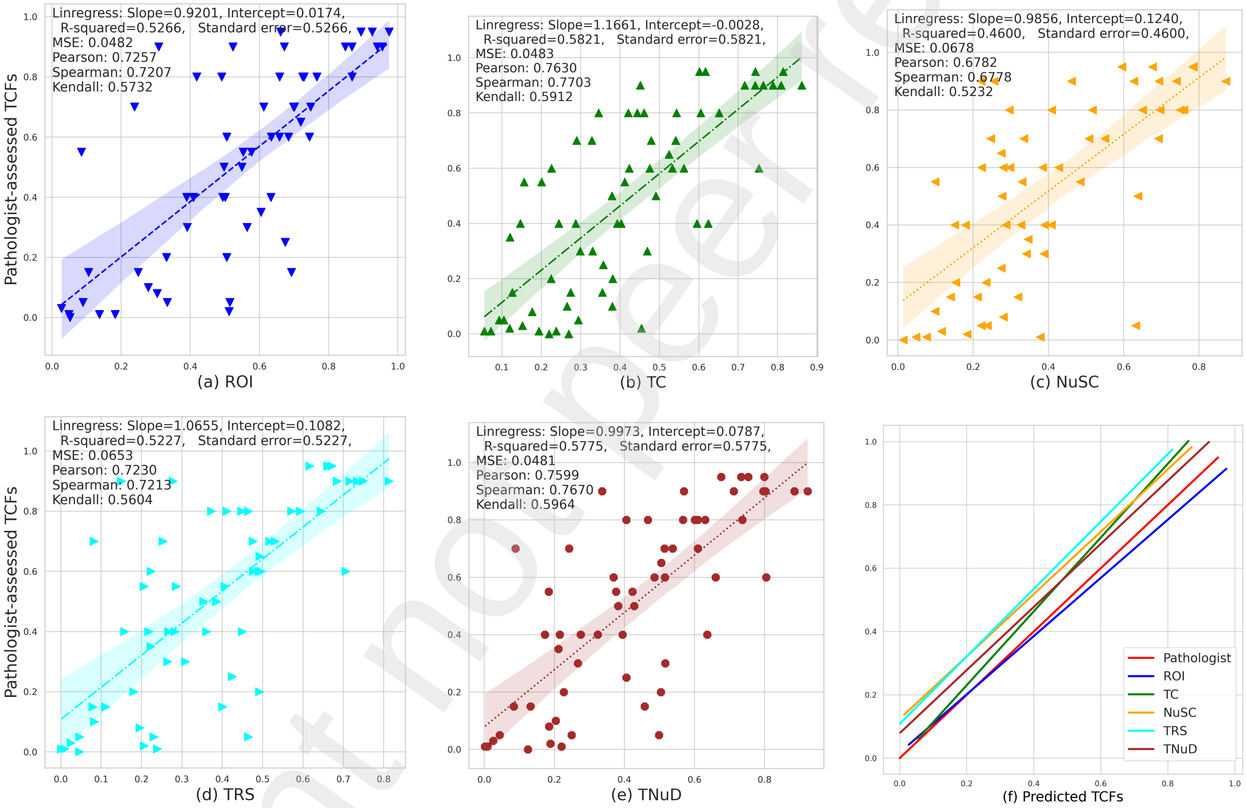
The TRS-based method displays similar shortcomings, with a higher MSE (0.0653) and relatively weaker correlations (Pearson: 0.7230, Spearman: 0.7213). Like other breast cancer-trained models, the performance of the TRS-based method on ovarian cancer tissue is compromised due to organ-specific differences. The suboptimal segmentation of tissue regions in ovarian cancer slides highlights the challenges of transferring models trained on one cancer type to another, emphasizing the need for organ-specific datasets.

Similarly, the NuSC-based method shows the highest MSE (0.0678) among all evaluated methods, indicating substantial errors in TCF prediction. Its lower correlation coefficients (Pearson: 0.6782, Spearman: 0.6778) further highlight its inability to provide reliable and consistent TCF estimates for ovarian cancer. This performance issue stems from the lack of NuSC datasets specifically for ovarian cancer. We can conclude that the NuSC-based model, trained exclusively on breast cancer data, struggles to accurately identify cells and classify their types in ovarian cancer images.

In summary, all evaluated methods were trained exclusively

Table 6: Comparison of the correlation between the predicted TCF of different estimation methods and pathologists' annotations on ovarian cancer.

Model	MSE↓	Pearson↑	Spearma↑	Linear			
				slope	intercept	R-values ↑	Std Err ↓
ROI	0.0482	0.7257	0.7207	0.9201	0.0174	0.5266	0.1146
TC	0.0483	**0.7630**	**0.7703**	1.1661	-0.0028	**0.5821**	0.1225
NuSC	0.0678	0.6782	0.6778	0.9856	0.1240	0.4600	0.1402
TRS	0.0653	0.7230	0.7213	1.0655	0.1082	0.5227	0.1337
TNuD	**0.0481**	0.7599	0.7670	0.9973	0.0787	0.5775	**0.1120**



(a) ROI (b) TC (c) NuSC (d) TRS (e) TNuD (f) Predicted TCFs

Figure 6: The visualization of the linear regression model between TCF predicted by different methods and pathologists' annotations on ovarian cancer.

on breast cancer data, limiting their predictive accuracy when applied to ovarian cancer. Among them, the TNuD-based workflow shows the most balanced performance, achieving relatively stable, unbiased predictions. Although the TC-based method achieves slightly stronger correlation metrics, it suffers from greater variability. Other methods, including ROI-based, TRS-based, and NuSC-based, exhibit more pronounced limitations. Overall, the results highlight the TNuD-based method's promising transferability and suggest that further ovarian-specific optimization is needed to enhance performance.

Discussion

Experimental results demonstrate that our TNuD-based approach exhibits superior performance in precision and reliability, with strong alignment to pathologist annotations, confirming its suitability for rapid, large-scale tumour-burden assessment in clinical applications. Across two breast-cancer test cohorts, the workflow reduced the median MSE by 25-40 % and increased the Pearson correlation by up to 0.08 relative to the best single-scale baseline, while maintaining minute-level inference time on a single GPU. The model trained on breast-cancer data also shows promising transferability to ovarian-cancer WSIs, suggesting cross-organ robustness.

From a theoretical perspective, our findings confirm that explicitly coupling macro-scale tissue architecture with micro-scale nuclear composition yields a synergistic gain unattainable for region-only or nucleus-only pipelines in our experiments. The density-matrix formalism introduced herein provides a principled means to encode spatial priors and can be extended to other histological attributes, such as stromal subtypes or immune foci. Compared with ROI-based and TC-based workflows–both of which misidentify mixed patches as tumor–we observed a markedly lower over-estimation slope on breast slides rich in lymphocytic infiltrates. This finding indicates that multi-scale fusion not only aggregates complementary information but also mitigates scale-specific biases.

Clinically, our method offers pathologists and oncologists a fast, accurate, and interpretable tool for evaluating tumour purity. High-confidence TCF scores can be used to triage low-purity samples before costly genomic assays, to adjust variant-calling thresholds, or to stratify patients in trials where tumour purity is a prognostic factor (e.g. HER2-, EGFR-, or PD-L1-targeted therapies). The framework's intermediate maps further support visual audit and facilitate integration into routine sign-out workflows.

Although the results affirm the clinical promise of the TNuD-based method, several aspects remain to be strengthened. Even with a relatively small sample size for breast cancer, our method produced statistically significant improvements, highlighting the robustness of its multi-scale integration approach. Nevertheless, the limited dataset size inherently restricts generalisability, and performance could likely be further enhanced with access to larger, diverse datasets. Furthermore, variability in imaging quality—including differences in staining protocols, slide preparation, and scanner settings—may influence performance when models are applied across clinical centres. In addition, full clinical translation will require regulatory approval, workflow integration, and clinician training, all of which lie beyond the present scope. Future research incorporating larger, more diverse datasets and addressing these practical considerations is needed to further enhance the model's robustness and applicability.

In conclusion, this work makes significant progress in advancing TCF estimation by addressing key limitations related to speed, accuracy, and interpretability. This comprehensive, multiscale TNuD-based approach more effectively captures the heterogeneity of tumor tissues, underscoring the importance of integrative methods in computational pathology. It offers higher precision and is therefore well-suited for clinical implementation where high accuracy is essential. Additionally, it makes detailed tissue region and nuclear feature information readily available to pathologists and clinicians, supporting accurate clinical decision-making. Building on these findings, future work will explore cross-cancer fine-tuning, multi-centre stain normalisation, and the fusion of TNuD features with genomic copy-number profiles or spatial-transcriptomic maps. Prospective, multi-institutional trials and an interactive visual interface are also planned to expedite regulatory approval and real-time adoption in diagnostic practice. Collectively, these efforts will further strengthen computational pathology and ultimately support more accurate cancer diagnosis and treatment decisions.

CRediT authorship contribution statement

Lulu Qin: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review and editing, Project administration. **Xiao Yang**: Conceptualization, Methodology, Funding acquisition, Resources, Supervision, Writing – review and editing. **Zhigang Pei**: Data curation, Formal analysis, Validation, Supervision, Writing – review and editing. **Susan Fotheringham**: Supervision, Writing – review and editing. **Xianhong Xu**: Conceptualization, Methodology, Supervision, Writing – review and editing. **Zexuan Zhu**: Conceptualization, Methodology, Funding acquisition, Resources, Supervision, Writing – review and editing. All authors reviewed and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We sincerely thank Prof David J. Kerr, University of Oxford, Oxford, UK, for his valuable feedback on our manuscript.

Funding

This work was supported in part by the National Key Research and Development Program of China, under Grant 2022YFF1202104, and in part by the National Natural Science Foundation of China under Grant 62471310.

References

Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M., et al., 2019. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics 35, 3461–3467.

Anderson, N.M., Simon, M.C., 2020. The tumor microenvironment. Current Biology 30, R921–R925.

Augustine, T.N., 2022. Weakly-supervised deep learning models in computational pathology. eBioMedicine 81, 104117.

Bejarano, L., Jordão, M.J., Joyce, J.A., 2021. Therapeutic targeting of the tumor microenvironment. Cancer discovery 11, 933–959.

Benelli, M., Romagnoli, D., Demichelis, F., 2018. Tumor purity quantification by clonal DNA methylation signatures. Bioinformatics 34, 1642–1649.

Brendel, M., Getseva, V., Al Assaad, M., Sigouros, M., Sigaras, A., Kane, T., Khosravi, P., Mosquera, J.M., Elemento, O., Hajirasouliha, I., 2022. Weakly-supervised tumor purity prediction from frozen h&e stained slides. EBioMedicine 80.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), Munich, Germany. pp. 801–818.

Cheng, H., Liu, X., Zhang, J., Dong, X., Ma, X., Zhang, Y., Meng, H., Chen, X., Yue, G., Li, Y., et al., 2025. Glmkd: Joint global and local mutual knowledge distillation for weakly supervised lesion segmentation in histopathology images. Expert Systems with Applications 279, 127425.

Cheng, J., He, J., Wang, S., Zhao, Z., Yan, H., Guan, Q., Li, J., Guo, Z., Ao, L., 2020. Biased Influences of Low Tumor Purity on Mutation Detection in Cancer. Frontiers in Molecular Biosciences 7.

Choi, S., Cho, S.I., Jung, W., Lee, T., Choi, S.J., Song, S., Park, G., Park, S., Ma, M., Pereira, S., Yoo, D., Shin, S., Ock, C.Y., Kim, S., 2023. Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer. npj Breast Cancer 9, 1–13.

Crosby, D., Bhatia, S., Brindle, K.M., Coussens, L.M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R.C., Gambhir, S.S., Kuhn, P., et al., 2022. Early detection of cancer. Science 375, eaay9040.

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M., . Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer 1, 800–810.

Gerardin, Y., Shenker, D., Hipp, J., Harguindeguy, N., Juyal, D., Shah, C., Javed, S.A., Thibault, M., Nercessian, M., Sanghavi, D., et al., 2024. Foundation ai models predict molecular measurements of tumor purity. Cancer Research 84, 7402–7402.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. 58, 101563.

Haider, S., Tyekucheva, S., Prandi, D., Fox, N.S., Ahn, J., Xu, A.W., Pantazi, A., Park, P.J., Laird, P.W., Sander, C., et al., 2020. Systematic assessment of tumor purity and its clinical implications. JCO precision oncology 4, 995–1005.

Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., et al., 2024. Cellvit: Vision transformers for precise cell segmentation and classification. Medical Image Analysis 94, 103143.

Jeon, H.W., Kim, Y.D., Sim, S.B., Moon, M.H., 2022. Predicting prognosis using a pathological tumor cell proportion in stage i lung adenocarcinoma. Thoracic Cancer 13, 1525–1532.

Kang, L.I., Sarullo, K., Marsh, J.N., Lu, L., Khonde, P., Ma, C., Haritunians, T., Mujukian, A., Mengesha, E., McGovern, D.P.B., Stappenbeck, T.S., Swamidass, S.J., Liu, T.C., 2024. Development of a deep learning algorithm for Paneth cell density quantification for inflammatory bowel disease. eBioMedicine 110, 105440.

Koo, B., Rhee, J.K., 2021. Prediction of tumor purity from gene expression data using machine learning. Briefings in Bioinformatics 22, 1–9.

Liu, B., Li, C., Li, Z., Wang, D., Ren, X., Zhang, Z., 2020. An entropy-based metric for assessing the purity of single cell populations. Nature Communications 11, 3155.

Liu, S., Amgad, M., More, D., Rathore, M.A., Salgado, R., Cooper, L.A.D., 2024a. A panoptic segmentation dataset and deep-learning approach for explainable scoring of tumor-infiltrating lymphocytes. npj Breast Cancer 10, 1–10.

Liu, Y., Bai, X., Wang, J., Li, G., Li, J., Lv, Z., 2024b. Image semantic segmentation approach based on deeplabv3 plus network with an attention mechanism. Engineering Applications of Artificial Intelligence 127, 107260.

Locallo, A., Prandi, D., Fedrizzi, T., Demichelis, F., 2019. TPES: Tumor purity estimation from SNVs. Bioinformatics 35, 4433–4435.

Menzel, M., Endris, V., Schwab, C., Kluck, K., Neumann, O., Beck, S., Ball, M., Schaaf, C., Fröhling, S., Lichtner, P., Schirmacher, P., Kazdal, D., Stenzinger, A., Budczies, J., 2023. Accurate tumor purity determination is critical for the analysis of homologous recombination deficiency (HRD). Translational Oncology 35, 101706.

Oner, M.U., Chen, J., Revkov, E., James, A., Heng, S.Y., Kaya, A.N., Alvarez, J.J.S., Takano, A., Cheng, X.M., Lim, T.K.H., Tan, D.S.W., Zhai, W., Skanderup, A.J., Sung, W.K., Lee, H.K., 2022. Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. Patterns 3.

Osinski, B.L., BenTaieb, A., Ho, I., Jones, R.D., Joshi, R.P., Westley, A., Carlson, M., Willis, C., Schleicher, L., Mahon, B.M., Stumpe, M.C., 2022. Artificial intelligence-augmented histopathologic review using image analysis to optimize DNA yield from formalin-fixed paraffin-embedded slides. Modern Pathology 35, 1791–1803.

Priego-Torres, B.M., Lobato-Delgado, B., Atienza-Cuevas, L., Sanchez-Morillo, D., 2022. Deep learning-based instance segmentation for the precise automated quantification of digital breast cancer immunohistochemistry images. Expert Systems with Applications 193, 116471.

Revkov, E., Kulshrestha, T., Sung, K.W.K., Skanderup, A.J., 2023. PUREE: Accurate pan-cancer tumor purity estimation from gene expression data. Communications Biology 6, 1–10.

Sakamoto, T., Furukawa, T., Pham, H.H.N., Kuroda, K., Tabata, K., Kashima, Y., Okoshi, E.N., Morimoto, S., Bychkov, A., Fukuoka, J., 2022. A collaborative workflow between pathologists and deep learning for the evaluation of tumour cellularity in lung adenocarcinoma. Histopathology 81, 758–769.

Schoenpflug, L.A., Chatzipli, A., Sirinukunwattana, K., Richman, S., Blake, A., Robineau, J., Mertz, K.D., Verrill, C., Leedham, S.J., Hardy, C., Whalley, C., Redmond, K., Dunne, P., Walker, S., Beggs, A.D., McDermott, U., Murray, G.I., Samuel, L.M., Seymour, M., Tomlinson, I., Quirke, P., Consortium, S., Rittscher, J., Maughan, T., Domingo, E., Koelzer, V.H., 2025. Tumour purity assessment with deep learning in colorectal cancer and impact on molecular analysis. The Journal of Pathology 265, 184–197.

Shi, X., Wang, X., Yao, W., Shi, D., Shao, X., Lu, Z., Chai, Y., Song, J., Tang, W., Wang, X., 2024. Mechanism insights and therapeutic intervention of tumor metastasis: latest developments and perspectives. Signal Transduction and Targeted Therapy 9, 192.

Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A., 2023. Cancer statistics, 2023. CA: a cancer journal for clinicians 73.

Silva, A.B., Martins, A.S., Tosta, T.A.A., Neves, L.A., Servato, J.P.S., de Araújo, M.S., de Faria, P.R., do Nascimento, M.Z., 2022. Computational analysis of histological images from hematoxylin and eosin-stained oral epithelial dysplasia tissue sections. Expert Systems with Applications 193, 116456.

Su, A., Lee, H., Tan, X., Suarez, C.J., Andor, N., Nguyen, Q., Ji, H.P., 2022. A deep learning model for molecular label transfer that enables cancer cell identification from histopathology images. NPJ precision oncology 6, 14.

Sun, P., He, J., Chao, X., Chen, K., Xu, Y., Huang, Q., Yun, J., Li, M., Luo, R., Kuang, J., et al., 2021. A computational tumor-infiltrating lymphocyte assessment method comparable with visual reporting guidelines for triple-negative breast cancer. EBioMedicine 70.

Tan, M., Le, Q.V., 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint arXiv:1905.11946 .

Tbeileh, N., Timmerman, L., Mattis, A.N., Toriguchi, K., Kasai, Y., Corvera, C., Nakakura, E., Hirose, K., Donner, D.B., Warren, R.S., Karelehto, E., 2023. Metastatic colorectal adenocarcinoma tumor purity assessment from whole exome sequencing data. PLOS ONE 18, e0271354.

Vykopal, I., Hudec, L., Kveton, M., Fabian, O., Felsoova, A., Benesova, W., 2023. Deeplabv3+ ensemble for diagnosis of cardiac transplant rejection, in: Computer Vision Systems (ICVS), Springer Nature Switzerland. pp. 112–122.

Wang, J., Qin, L., Chen, D., Wang, J., Han, B.W., Zhu, Z., Qiao, G., 2023. An improved Hover-net for nuclear segmentation and classification in histopathology images. Neural Computing and Applications 35, 14403–14417.

Wei, N., Zhu, H., Li, C., Zheng, X., 2021. Purimeth: An integrated web-based tool for estimating and accounting for tumor purity in cancer DNA methylation studies. Mathematical Biosciences and Engineering 18, 8951–8961.

West, N., Dattani, M., McShane, P., Hutchins, G., Grabsch, J., Mueller, W., Treanor, D., Quirke, P., Grabsch, H., 2010. The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients. British journal of cancer 102, 1519–1523.

Yu, T., Huang, Q., Zhao, X., Zhang, S., Zhang, Q., Fan, X., Liu, G., 2023. Tumour purity as an underlying key factor in tumour mutation detection in colorectal cancer. Clinical and Translational Medicine 13.

Zhang, X., Venkatachalapathy, S., Paysan, D., Schaerer, P., Tripodo, C., Uhler, C., Shivashankar, G., 2024. Unsupervised representation learning of chromatin images identifies changes in cell state and tissue organization in dcis. Nature Communications 15, 6112.

Zheng, X., Zhao, Q., Wu, H.J., Li, W., Wang, H., Meyer, C.A., Qin, Q.A., Xu, H., Zang, C., Jiang, P., Li, F., Hou, Y., He, J., Wang, J., Wang, J., Zhang, P., Zhang, Y., Liu, X.S., 2014. MethylPurify: Tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome Biology 15, 419.