Title: Normal breast tissue classifiers assess large-scale tissue compartments with high accuracy

Siyuan Chen[1†], Mario Parreno-Centeno[1†], Graham Booker[1], Gregory Verghese[1,2], Fathima Sumayya Mohamed[1], Salim Arslan[3], Pahini Pandya[3], Aasiyah Oozeer[4], Marcello D'Angelo[4], Rachel Barrow[5], Rachel Nelan[5], Marcelo Sobral-Leite[6], Fabio de Martino[7], Cathrin Brisken[7], Mathew J. Smalley[8], Esther Lips[6], Cheryl Gillett[4], Louise J. Jones[5], Christopher R.S. Banerji[1,9,10], Sarah E. Pinder[11], Anita Grigoriadis[1,2,*]

Authors affiliations

[1] Cancer Bioinformatics, School of Cancer & Pharmaceutical Sciences, Faculty of Life Sciences and Medicine, King's College London, London, UK.

[2] The Breast Cancer Now Research Unit, School of Cancer and Pharmaceutical Sciences, Faculty of Life Sciences and Medicine, King's College London, London, UK.

[3] Panakeia Technologies, London, UK.

[4] King's Health Partners Cancer Biobank, King's College London, London, UK.

[5] Centre for Tumour Biology, Barts Cancer Institute, Queen Mary University of London, London, UK.

[6] Division of Molecular Pathology, Netherlands Cancer Institute, Amsterdam, the Netherlands.

[7] Swiss Institute for Experimental Cancer Research, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

[8] European Cancer Stem Cell Research Institute, School of Biosciences, Cardiff University, Cardiff, UK.

[9] The Alan Turing Institute, London, UK.

[10] University College London NHS Trust, London, UK.

[11] Research Oncology, Division of Cancer Studies, King's College London, Guy's Hospital, London, UK.

† These authors contributed equally

* **Corresponding author**

Professor Anita Grigoriadis, Cancer Bioinformatics, School of Cancer and Pharmaceutical Sciences, Faculty of Life Sciences and Medicine, King's College London, Innovation Hub, Cancer Centre at Guy's Hospital, London SE1 9RT, UK, (+44) 20 7188 2360, anita.grigoriadis@kcl.ac.uk

Word count: 4061 words without methods

Keywords: normal breast tissue, tissue type classification, convolutional neural network (CNN)

Summary (227/300 words)

Background Cancer research emphasises early detection, yet quantitative methods for analysing normal tissue remain limited. Hematoxylin and eosin (H&E)-stained tissues in digitised whole slide images (WSIs) enable computational histopathology; however, artificial intelligence (AI)-based analyses of normal breast tissue (NBT) remain scarce.

Methods We curated 70 WSIs of NBTs from multiple sources with pathologist-guided manual annotations of epithelium, stroma, and adipocytes, and developed robust convolutional neural network (CNN)-based, patch-level classification models, named *NBT-Classifiers*, to tessellate and classify NBTs at different scales. Data and code are available at https://github.com/cancerbioinformatics/OASIS and https://github.com/cancerbioinformatics/NBT-Classifier.

Findings Across three external cohorts, *NBT-Classifiers* trained on 128 x 128 µm and 256 x 256 µm patches achieved AUCs of 0·98–1·00. Two explainable artificial-intelligence (AI)-visualisation techniques confirmed the biological relevance of tissue class predictions. An end-to-end WSI pre-processing framework was then integrated, capable of localising lobules and peri-lobular stroma. The outputs are compatible with the QuPath v0.3.0 platform and enable downstream digital image analysis, such as texture and nuclei morphology assessment, at the patch level.

Interpretation *NBT-Classifiers* represent a robust, generalisable, end-to-end deep learning framework, which will enable broader application in studies of normal tissues, in the context of breast and breast cancer.

Funding The Breast Cancer Research Trust, Breast Cancer Now (and their legacy charity Breakthrough Breast Cancer), the Medical Research Council (MRC) [MR/X012476/1], Cancer Research UK [CRUK/07/012, KCL-BCN-Q3], and CRUK City of London Centre Award [CTRQQR-2021/100004].

Research in context

Evidence before this study

We searched PubMed from Jan 1, 2013, to Mar 20, 2025, using the query (("normal breast" OR "terminal duct lobular unit" OR "TDLU" OR "acini") AND ("artificial intelligence" OR "deep learning" OR "convolutional neural network" OR "CNN" OR "machine learning" OR "automation" OR "classification" OR "segmentation" OR "detection")) AND (journal article[pt] NOT review[pt]) NOT ("computed tomography" OR "CT" OR "ultrasound" OR "sonography" OR "US" OR "mammography" OR "mammogram" OR "Raman" OR "single-cell" OR "MRI"). This yielded 264 initial hits and after a thorough manual review, only eight studies were found to be relevant for using deep learning or machine learning to automate tissue recognition or quantification of normal breast tissues (NBTs) on digitised histological images. All these methods are highly specialised for detecting and quantifying lobules or quantifying tissue composition, and none of them are specifically tailored for providing patch-level tissue classifications on WSIs of NBTs. Moreover, NBTs are heavily underrepresented in large annotated WSI databases. A recent literature review summarised publicly available breast hematoxylin and eosin (H&E) WSI datasets from 2015 to 2023, identifying 17 datasets comprising a total of 10,385 breast H&E WSIs. Of these, two datasets contain female normal breast WSIs with pathology-guided manual annotations.

Added value of this study

Motivated by the research gap, we collected a comprehensive WSI dataset that captures real-world variability present across NBTs and provided by expert ground-truth manual annotations. Based on extensive cross-validation and external testing, we developed robust patch-level classification models to classify three major tissue compartments within NBTs, at different scales.

Implications of all the available evidence

With high AUCs of 0·98–1·00 across three external cohorts, our artificial-intelligence (AI) tool can be used to automatically classify tissue compartments and localise lobular regions for downstream normal breast research. These approaches have the potential to enhance our understanding of how various NBT components contribute to both benign and malignant breast

pathology and lay the groundwork for the development of more advanced deep learning models and spatial-defined molecular large-scale analyses in the future.

Background

Normal breast tissue (NBT) research is gaining momentum for the early detection of breast cancer.[1] The epithelium, organised into ducts and lobules, also referred to as terminal duct-lobular units (TDLUs), is of primary importance, as these structures are where most physiological processes and pathological alterations occur.[2] Apart from epithelium, recent evidence suggests that the stroma, particularly the peri-lobular stroma, may also harbour precursor alterations and have a crucial role in breast cancer progression.[3] Therefore, in-depth histopathological research into these structures, especially in women with varying breast cancer risk, holds promise for detecting early signs of cancer initiation in breast tissues that appear "normal". This is particularly significant for women at higher breast cancer risk, such as germline *BRCA1/2* mutation (*gBRCA1/2m*) carriers.[4]

Histopathology remains the gold standard for diagnosing cancer, relying on microscopic examination to detect tissue abnormalities. In routine practice, pathologists first survey the entire tissue section on a glass slide before focusing on pathological regions at higher magnification for a detailed assessment. Whole slide images (WSIs) enable the hierarchical storage of microscopic histological images scanned at multiple magnifications.[5] Their ease of management and ability to facilitate rapid sharing has led to their growing integration into clinical workflows,[6] as well as driving advancements in digital image analysis using platforms such as QuPath.[7] This shift has led to the recent growth of computational pathology (CPath),[8] a rapidly evolving field that leverages artificial intelligence (AI) and deep learning, particularly convolutional neural networks (CNNs) and the latest vision transformer (ViT)-based foundation models,[9] for high-throughput analysis of large-scale WSI datasets. Automatically classifying tissue and decomposing individual tissue compartments into smaller pieces, namely patches, are crucial WSI pre-processing steps to uncover clinically relevant histological features using CPath frameworks, such as multiple-instance learning (MIL).[10]

For tissue classification, numerous CNN-based models have been developed for breast cancer,[11] including approaches that classify tissue compartments within normal tissues adjacent to tumours, such as HistoROI.[12] By contrast, several studies have focused on "independent normal" breast tissues and developed methods for detecting and quantifying lobules[13-18] or assessing overall tissue composition.[19, 20] Nonetheless, none of these approaches were specifically designed to provide patch-level tissue classifications on digitised WSIs of NBTs. Moreover, all these studies typically relied on a single source of NBTs, which may limit the

generalisability of the resulting models. Generally, NBTs are heavily underrepresented in large annotated WSI databases. A recent literature review summarised publicly available breast hematoxylin and eosin (H&E) WSI datasets from 2015 to 2023,[21] identifying 17 datasets comprising a total of 10,385 breast H&E WSIs. Of these, only two datasets contain NBTs (~350 female normal breast WSIs), and just one dataset (~44 WSIs) includes manual annotations. Thus, there is a clear need for a universal patch-level classification model for NBTs, developed on a curated dataset that captures real-world variability and is supported by ground-truth manual annotations.

In this study, we present *NBT-Classifiers*, CNN-based deep learning models designed to identify patches of epithelium, stroma and adipocytes at two different scales. We provide pathologist-guided manually annotated NBTs on digitised WSIs, sourced from five cohorts, including NBTs from healthy women without high-risk germline mutations who underwent reduction surgery or core biopsy, *gBRCA1/2m* carriers who underwent a risk-reducing mastectomy, as well as contralateral and ipsilateral NBTs from breast cancer patients. *NBT-Classifiers* achieved robust performance across three external cohorts and cell type classification of the predictions was further validated by explainable AI-based approaches.[22, 23] Additionally, an end-to-end framework was integrated, outputting pseudo patch-level annotations of target regions, including lobules and their microenvironment. Collectively, our *NBT-Classifiers* provide a primary tool for studying NBTs at sub-tissue and cellular levels on large-scale digitised images.

Results

Optimisation of *NBT-Classifiers*

In practice, pixel-level annotated digitised histological image datasets of NBTs are notably scarce. Furthermore, the biological heterogeneity of NBTs from diverse sources is often overlooked when assembling such datasets. To address this, we curated a meticulously pathologist-guided annotated dataset of 70 WSIs of NBTs, sourced from healthy donors undergoing reduction mammoplasty or core biopsy (*non-gBRCA1/2m* carriers), *gBRCA1/2m* carriers undergoing risk-reducing mastectomy, and contralateral or ipsilateral normal tissues from breast cancer patients (**Fig. 1a**, **Supplementary Table 1**, **Methods**). These NBT WSIs were resourced from five distinct patient cohorts: the King's Health Partners Cancer Biobank

(KHP) in London (UK), the Netherlands Cancer Institute (NKI) in Amsterdam (Netherlands), the Barts Cancer Institute (BCI) in London (UK), the École Polytechnique Fédérale de Lausanne (EPFL) in Lausanne (Switzerland), and the publicly available Susan G. Komen Tissue Bank (SGK)[24] (**Fig. 1b**). Collectively, these digitised images capture both technical variability (differences in H&E-staining and scanning) and biological diversity (variations in tissue sources and a broad patient age range) within NBTs. The WSIs from the NKI and BCI cohorts were used for model training and optimisation, while those from the KHP, EPFL, and SGK cohorts were reserved for external validation. Expert manual annotations of epithelium, stroma, and adipocytes were performed under the supervision of a consultant pathologist (SP) using QuPath v0.3.0 (**Fig. 1c**, **1d**, **Supplementary Fig. 1**, **Methods**). The pathology-guided annotated WSIs of NBTs are available in the nOrmal breASt tISsue Dataset (OASIS) repository, accessible at https://github.com/cancerbioinformatics/OASIS.

NBTs inherently exhibit a hierarchical organisation. For instance, within the epithelium, structures ranging from individual epithelial cells, acini to lobules can be captured at fields of view (FOVs) of 64 x 64 µm, 128 x 128 µm, and 256 x 256 µm (**Supplementary Fig. 2**). Therefore, we aimed to train models enabling tissue classification at different scales to offer greater analytical flexibility for downstream analysis. To train these models, we generated patch-level datasets using QuPath v0.3.0 for the NKI and BCI cohorts. These datasets consisted of non-overlapping patches with sizes of 224 x 224, 512 x 512, and 1024 x 1024 pixels, all extracted at a resolution of 0·25 µm/pixel (**Supplementary Table 2**, **Methods**). In the training pipeline (**Fig. 1c**), patches were sampled in a class-balanced manner to ensure equal representation of tissue types and were subjected to stain normalisation to minimise cohort variability (**Supplementary Fig. 3**). The model employs a pre-trained CNN backbone for feature extraction (see Methods), with a trainable classification head that predicts the probabilities of epithelium, stroma, and adipocytes (**Supplementary Fig. 4**).

Given the critical role of stain normalisation and the selection of CNN backbone, we initially performed three-fold cross-validation experiments to optimise these parameters. The stain normalisation methods evaluated included Reinhard, Vahadane, Macenko, and StainGAN (**Supplementary Fig. 5**, **Methods**).[25] For the CNN backbones, we tested MobileNet, ResNet50, DenseNet, and InceptionV3, using weights pre-trained on ImageNet.[11] These experiments were conducted using patch-level datasets containing non-overlapping 512 x 512-pixel patches from the NKI and BCI cohorts. In each fold, 250 patches per tissue class were randomly sampled from each WSI (**Supplementary Fig. 6**), with patches split at the WSI level: 20 WSIs for training

(10 from each cohort), 4 WSIs for internal validation (2 from each cohort), and 8 WSIs for internal testing (4 from each cohort) (**Supplementary Fig. 7**). The combination of Reinhard stain normalisation and the MobileNet backbone yielded the best overall performance (**Fig. 1e**).

Then, configured with Reinhard stain normalisation and the MobileNet backbone (**Supplementary Fig. 7**), we trained models using non-overlapping patches of 224 x 224, 512 x 512, and 1024 x 1024 pixels from the NKI and BCI cohorts. Given the varying number of available patches, we randomly sampled 150 patches per tissue class from each WSI (**Supplementary Fig. 6**). The models trained on larger FOVs generally achieved higher accuracies (**Fig. 1f**), while the 224 x 224-pixel model exhibited the lowest accuracy and was more prone to misclassifying epithelium (**Supplementary Fig. 8**). Due to the suboptimal performance of the 224 x 224-pixel model, we proceeded to retrain two final versions of the *NBT-Classifier*: one using 512 x 512-pixel patches and the other using 1024 x 1024-pixel patches.

External evaluations of *NBT-Classifiers*

Next, we evaluated the optimised *NBT-Classifiers* on the external KHP, EPFL, and SGK cohorts (**Methods**). We first evaluated the 512px-based *NBT-Classifier* using the receiver operating characteristic (ROC) analysis. The class-specific area under the curve (AUC) values were 0·98 for epithelium, 0·98 for stroma, and 1·00 for adipocytes across all cohorts (**Fig. 2a**). Cohort-specific AUCs slightly varied, in the KHP cohort, the AUCs were 0·99 for epithelium, 1·00 for stroma, and 1·00 for adipocytes; in the EPFL cohort, AUCs were 0·97 for epithelium, 0·93 for stroma, and 0·99 for adipocytes; and in SGK cohort, AUCs were 1·00 for all tissue types (**Supplementary Fig. 9**). These findings underscore the discriminative power and generalisability of the 512px-based *NBT-Classifier.* The quality of feature representations provides an additional measure of a model's performance. We then visualised features output from the last layer (the layer before the output layer) of the 512px-based *NBT-Classifier* using t-distributed Stochastic Neighbor Embedding (t-SNE). Patches predicted as epithelium, stroma, and adipocytes formed well-separated visual clusters (**Fig. 2b**), though a notable domain shift was observed across cohorts (**Supplementary Fig. 10**). Moreover, the visual clusters aligned well with the actual histology of epithelium, stroma, and adipocytes (**Fig. 2c**, **Methods**). These results demonstrate that the model successfully learned meaningful representations that correlate with tissue histology. When comparing the predictions with ground-truth annotations, most misclassifications occurred at the boundaries between class-specific visual clusters where

histological features are less distinct (**Fig. 2d**). Correspondingly, the confidence of model's predictions, approximated by the probability associated with the predicted tissue class, was notably lower at these boundaries (**Fig. 2e**). This suggests that the misclassifications are not due to a bias in the model's innate judgement, but rather to ambiguity in the testing patches where tissue classes are less clearly defined.

To benchmark the performance of the *NBT-Classifier*, we implemented HistoROI (**Methods**), a CNN-based tissue classification model trained to distinguish epithelium, stroma, and adipocytes (referred to as adipose in the original study), along with lymphocytes, miscellaneous tissues, and artefacts. Although initially developed using WSIs from a single breast cancer cohort, HistoROI had demonstrated strong generalisability to unseen tissues. We focussed on the 1024px-based *NBT-Classifier*, as it had achieved similarly robust class-specific AUCs of 0·99 for epithelium, 0·98 for stroma, and 1·00 for adipocytes (**Fig. 2f**). HistoROI demonstrated comparable AUCs for epithelium (1·00) and stroma (0·98), but its performance was lower for adipocytes, with an AUC of 0·95 (**Supplementary Fig. 11**). In terms of cohort-specific AUCs, the two models showed similar performance in the KHP and SGK cohorts (**Supplementary Fig. 12**, **13**). In KHP, the 1024px-based *NBT-Classifier* and HistoROI both achieved perfect AUCs of 1·00 for epithelium and stroma, with a minor difference in adipocytes (1·00 for the *NBT-Classifier* vs. 0·96 for HistoROI). In SGK, the 1024px-based *NBT-Classifier* achieved 0·99 for epithelium, 0·99 for stroma, and 1·00 for adipocytes, while HistoROI achieved 1·00 for all classes. However, in the EPFL cohort, the 1024px-based *NBT-Classifier* demonstrated an advantage, particularly for adipocytes. Specifically, the 1024px-based *NBT-Classifier* achieved AUCs of 0·98 for epithelium, 0·96 for stroma, and 0·99 for adipocytes, while HistoROI's AUCs were 0·99 for epithelium, 0·95 for stroma, and 0·91 for adipocytes. We further compared the classification accuracy of the two models and found that the 1024px-based *NBT-Classifier* outperformed HistoROI. It correctly classified 87% of epithelium, 94% of stroma, and 96% of adipocytes patches, compared to 80%, 91%, and 72% for HistoROI, respectively (**Supplementary Fig. 14**). These findings highlight the superior performance of the 1024px-based *NBT-Classifier*, particularly in distinguishing adipocytes in the context of NBTs. For both models, t-SNE visualisations showed well-separated tissue class clusters, consistent with expected histological features (**Supplementary Fig. 15**, **16**). Misclassifications were primarily observed at the boundaries between tissue clusters, where patches exhibited ambiguous features. Additionally, domain shift was evident in feature distributions from both the 1024px-based *NBT-Classifier* and HistoROI, suggesting variations across cohorts.

Tissue class predictions must be robust to biological variations, including those arising from patient age and the source of NBTs. Therefore, we also evaluated whether the *NBT-Classifiers* exhibited any bias toward these potential confounders. For a fair comparison, we sampled 15 patients from a total of 38 patients across three external cohorts, ensuring balanced representation across three age groups, namely “premenopausal years” <45 years: 5 patients, “menopausal years” 45–55 years: 5 patients, “postmenopausal years” >55 years: 5 patients. We sampled 30 patients from a total of 32 patients across NKI and BCI cohorts to include equal representation across three NBT sources: reduction mammoplasty (10 patients), risk-reducing mastectomy (10 patients), and NBTs from breast cancer patients (10 patients). To mitigate class imbalance, 500 patches per tissue class per WSI were randomly sampled when using the 512px-based model and 250 patches per class per WSI were sampled when using the 1024px-based model. These experiments were repeated three times. For both 512px- and 1024px-based *NBT-Classifiers*, we observed no bias in the accuracy concerning patient age in external cohorts or source of the NBT (**Supplementary Fig. 17**). Collectively, these results indicate the strong external performance of the *NBT-Classifiers.*

Model interpretability

To enhance the transparency of the *NBT-Classifiers’* decision-making process, we visualised high-attention regions in the model employed to make class-specific predictions using class activation mapping (CAM)[22] and gradient-weighted class activation mapping (Grad-CAM)[23]. Based on their distinct mechanisms (**Supplementary Fig. 18**), CAM typically generates more localised heatmaps that precisely highlight areas relevant to the final classification. In contrast, Grad-CAM may produce more diffuse or generalised heatmaps in some instances, due to its reliance on gradients to capture the model's learning dynamics. We first visually compared the CAM and Grad-CAM heatmaps for high-confidence patches of the predicted tissue class. Both methods consistently highlighted similar regions, capturing semantically meaningful patterns, such as cellular contents, collagen fibres, and adipocytes membranes, which are characteristic of epithelium, stroma, and adipocytes, respectively (**Fig. 3a**, **Supplementary Fig. 19**). When CAM was applied to a larger, representative region (5120 x 5120 pixels, 0·25 µm/pixel) encompassing all three tissue compartments, the previously observed patterns were reliably captured across non-overlapping patches, maintaining spatial continuity (**Fig. 3b**). When analysing the 1024px-based *NBT-Classifier*, the activation heatmaps even more prominently revealed the distinct histological structures that the model attended to for predicting each class,

namely cellular contents, non-cellular collagen fibres, and adipocytes membranes (**Fig. 3c**, **Supplementary Fig. 19**). These results further substantiate the interpretability of *NBT-Classifiers* and confirm that biologically meaningful histological features were employed by *NBT-Classifiers* during the decision-making process. Moreover, the patterns identified through CAM were reproducible across all cohorts (**Supplementary Fig. 20**), including images scanned at 20x magnification from the SGK cohort (**Supplementary Fig. 21**). We then applied CAM and Grad-CAM to assess patches with low-confidence predictions. The resulting heatmaps for all three tissue classes showed that the model's attention was proportionally distributed across mixed tissue regions (**Supplementary Fig. 22**). This analysis further suggests that the heterogeneity of tissue components within a patch contributes to the model's lower confidence in making predictions.

Large-scale NBT classification and visualisation

NBT-Classifiers are patch-level tissue classification models that, once trained on annotated patches, can be applied to entire WSIs. We then performed whole slide tissue classification for a WSI example using both the 512px- and 1024px-based *NBT-Classifiers*, with the results converted and imported into QuPath v0.3.0 for interactive histological inspection (**Fig. 4a**, **4b**, **Supplementary Fig. 23**, **Methods**). The heatmaps generated by both models showed strong alignment between the predicted tissue classes and the underlying histology, at different scales. For the same example, we further examined the spatial distribution of low-confidence tissue regions predicted by the 512px-based *NBT-Classifier*. Patches with confidence above 0·7 covered most tissue regions, while those with scores above 0·99 covered only partial regions, with gaps primarily at the boundaries between tissue compartments and within large lobules (**Fig. 4c**). These low-confidence patches in large lobules may explain the reduced accuracy of the 224px-based *NBT-Classifier*. The smaller patch size tends to capture more intra-lobular stroma, which leads to the misclassification of epithelium as stroma. Among all patches used for external validation, 98·2% had probabilities greater than 0·7, and 84% had probabilities greater than 0·99 (**Fig. 4d**). This analysis suggests that model confidence can serve as a valuable metric for capturing spatial and histological variations within each tissue compartment, making it particularly useful for patch selection in WSI analysis, such as in MIL.

For the representative larger tissue region (5120 x 5120 pixels, 0·25 µm/pixel), we generated a smoother heatmap using predicted tissue class probabilities (**Methods**). The high-resolution heatmap effectively delineated individual lobules and segmented tissue compartments with

smooth transitions (**Fig. 4e**). Additionally, we analysed six other regions from diverse patient age groups, all demonstrating similarly strong performance (**Supplementary Fig. 24**). These results highlight the potential of this approach to serve as a foundation for more advanced downstream image analyses, such as lobule detection.

End-to-end *NBT-classifier* framework

Building on the demonstrated utility of the proposed *NBT-Classifiers*, we integrated the models into an end-to-end *NBT-classifier* framework for WSI pre-processing (**Fig. 5a**). This pipeline generates tissue classification results that facilitate large-scale digital image analysis within target sub-tissue regions, such as lobules and their microenvironment, and support the patch selection process for more advanced deep learning frameworks, such as MIL.

The pipeline begins with HistoQC[26] to detect foreground tissue regions. Once identified, the region is tessellated into non-overlapping patches, which are then analysed by the corresponding *NBT-Classifier* to predict tissue classes. This produces a whole slide tissue class heatmap, from which lobules and their peri-lobular regions within a specified range can be detected and localised (**Fig. 5b**, **Methods**). For frameworks such as MIL, patches from target tissue regions can be selectively balanced and treated as a hyperparameter to optimise performance during training (**Fig. 5a**). Besides, the pipeline outputs a binary mask that localises individual lobules on each WSI and can be directly imported into QuPath v0.3.0 as pseudo lobule annotations (**Fig. 5c**). Within each lobule, QuPath v0.3.0's built-in nuclei detection and advanced spatial analysis tools, such as Delaunay triangulation, can be utilised for object-level lobule quantification. Additionally, the original patch-level classifications of the target regions (lobules and peri-lobular areas) can be exported and loaded into QuPath v0.3.0 to perform digital image analyses, such as texture analysis, across the slide at the patch level (**Fig. 5d**). These approaches build upon the previous efforts on standardised quantifications of TDLUs,[13-17] aiming at expanding pathomics measurements of lobules as well as their direct microenvironment, which can unlock novel biomarkers that are indicative of breast cancer precursors. The pipeline performed robustly across WSIs with varying staining and is compatible with both 40x and 20x magnifications (**Supplementary Fig. 25**), offering flexibility for studies involving heterogeneous datasets. The source code running this pipeline is available at: https://github.com/cancerbioinformatics/NBT-Classifier.

Discussion

AI-driven WSI analysis has given rise to groundbreaking advancements in histopathology.[8] In mammary gland research, while the research frontier remains largely focused on breast cancer, histologically normal tissues are significantly underrepresented among millions of digitised WSIs, with few applications specifically designed for facilitating in depth morphological characterisation of NBTs.[13-20] To address this gap, we compiled high-quality, manually annotated WSIs of NBTs, sourced from women with varying breast cancer risk across multiple institutions. Through extensive optimisation and external validation, we developed robust *NBT-Classifiers* capable of processing WSIs at different scales and performing patch-level tissue classification of epithelium, stroma, and adipocytes. Building on this, we integrated an end-to-end *NBT-Classifier* framework to ensure objective and reproducible WSI tessellation and patch selection from biologically important regions, such as lobules and their microenvironment. The outputs are compatible with QuPath v0.3.0, one of the most widely used digital image analysis platforms and can be seamlessly integrated with deep learning-based CPath frameworks such as MIL. Our models (https://github.com/cancerbioinformatics/NBT-Classifier) and datasets (https://github.com/cancerbioinformatics/OASIS) are publicly available, whereby the latter will be highly valuable resource for training pathology foundation models as they provide rich histology data of normal breast.

Currently, most AI applications in CPath are focused on lesions, with few developed for the analysis of normal tissues. One notable study trained a model using only normal skin WSIs, however, it did not focus on tissue type classification.[27] To the best of our knowledge, *NBT-Classifiers* represent the first deep learning-based CPath models specifically trained to classify tissue compartments within normal tissue without any lesions, in the context of the breast.

To establish fundamental tissue classification tools for NBTs, particularly as a primary step for downstream analyses, ensuring generalisability, robustness, and adaptability is essential. Achieving this requires a diverse and representative WSI dataset, ground-truth manual annotations, and seamless compatibility with downstream analytical frameworks. Previous NBT studies have typically relied on a single data source, such as benign breast biopsies,[13-15, 19] healthy donors,[16, 18] or reduction mammoplasty.[17] Only one exception incorporated NBTs from both prophylactic mastectomy and reduction mammoplasty (*gBRCA1/2m* status unknown).[20] However, these datasets, sourced from limited origins, do not capture the full diversity of NBTs, which may exhibit varying predispositions to cancer initiation and distinct histological

manifestations. In contrast, a key strength of our work is the incorporation of a broad spectrum of NBT variations, including differences in patient age, breast cancer risk, H&E staining protocols, and scanning platforms.

In the generation of AI, manual annotations should not be overlooked when developing supervised applications, especially in highly expert histopathological domains.[28] Our fully pathologist-supervised ground-truth likely explains the slightly better performance in adipocytes classification compared to HistoROI, which employed human-in-the-loop annotations. Among the previous NBT research, three studies provide manual annotations, yet their sample sizes and annotation levels vary (**Supplementary Table 3**). We expanded on this by providing detailed annotation of lobules, stroma and adipocytes that displayed biological variations.

Performing patch-level analysis within key regions of NBTs, such as lobules and their microenvironment, is essential for detecting early cancer precursors. Prior NBT studies have focused on well-established cancer risk biomarkers, such as TDLU involution,[13-16] immune infiltration in lobular regions,[17] alterations in tissue composition,[18, 19] and fibroglandular breast density.[20] Consequently, existing models and methods are highly specialised for detecting and quantifying lobules[13-18] or assessing tissue composition.[19, 20] In addition to these visible pathological biomarkers, recent studies have highlighted even greater predictive potential within NBTs at the single-cell level.[29] Given the extensive validation of genotype-phenotype links in various CPath studies focused on cancer,[10] analysing sub-tissue regions at the cellular and sub-cellular level shows significant promise for NBTs as well. In this context, the ability to adapt to broader downstream pathomics feature extraction tools, such as QuPath-based digital image analysis, or deep learning-driven CPath frameworks, opens new opportunities for evaluating high-throughput histopathological measurements. This, in turn, provides insights into localised histological markers that may signal the early stages of breast cancer initiation. Building upon this, advanced techniques such as spatial transcriptomics and spatial proteomics could uncover the underlying molecular mechanisms driving these histological changes. By linking molecular data with histological features, a deeper understanding of how early molecular alterations contribute to the development of breast cancer can be achieved.

One limitation of this study is the distribution of cohorts between the training and external validation sets. The NKI and BCI cohorts were selected for training due to their greater diversity in sources of NBTs. However, the external validation cohorts consisted of less diverse sources. As a result, models' performance may not fully reflect their ability to generalise across the full

spectrum of variations across NBTs, such as contralateral tissues or those from individuals undergoing risk-reducing mastectomy. Future studies should aim to validate the models on a more diverse set of NBTs to ensure broader applicability. Another limitation is the scope of the annotations, as only a subset of tissue types within the NBTs was included. Consequently, when applying the *NBT-Classifiers* to the whole WSI, some stroma patches were misclassified as epithelium. This misclassification occurred because regions such as artefacts, necrotic tissue, and blood vessels were not annotated, causing them to be incorrectly classified as epithelium. Incorporating these regions as additional classes could further improve the accuracy of whole slide tissue classification. While expert annotation remains essential due to the histological complexity of NBTs, strategies to improve efficiency, such as AI-powered annotation platforms or carefully guided crowdsourcing approaches,[28] could help scale the process. Biological assays or molecular markers[17] could further support the generation of more precise, cell-level ground-truth annotations. In this study, ViT-based foundation models were not used, as the near-perfect AUCs achieved suggest that the semantic distinctions between epithelium, stroma, and adipocytes were sufficiently pronounced to be captured by conventional CNN-based models. Nonetheless, future work could explore whether transformer-based backbones offer further performance gains or improved generalisability. Lastly, stain normalisation techniques, especially when applied across heterogeneous cohorts, play a critical role in improving the model's generalisability.

In summary, we present deep learning-based *NBT-Classifiers* for patch-level tissue type classification in NBTs. These classifications have the potential to enhance our understanding of how various NBT components contribute to both benign and malignant breast pathology and lay the groundwork for the development of more advanced deep learning models and spatial-defined molecular large-scale analyses in the future.

Methods (1862 words)

Patient cohorts

We collected in total 70 digitised formalin-fixed, paraffin-embedded (FFPE), H&E-stained images (n=70 patients) of NBTs from various sources, including core biopsies of healthy donors (non-*gBRCA1/2m* carriers, n=12 WSIs), women undergoing reduction mammoplasties (non-

gBRCA1/2m carriers, n=30 WSIs), *gBRCA1/2m* carriers undergoing risk-reducing mastectomy (n=21 WSIs), and contralateral or ipsilateral normal tissues from breast cancer patients (n=7 WSIs). A unique WSI was included per patient from the following five cohorts: the King's Health Partners Cancer Biobank (KHP) in London (UK) (n=16 WSIs), the Netherlands Cancer Institute (NKI) in Amsterdam (Netherlands) (n=16 WSIs), the Barts Cancer Institute (BCI) in London (UK) (n=16 WSIs), the École Polytechnique Fédérale de Lausanne (EPFL) in Lausanne (Switzerland) (n=10 WSIs), and the public Susan G. Komen Tissue Bank (SGK) (n=12 WSIs).[24] Patient ages ranged from 16 to 74 years. WSIs from the KHP, BCI, and EPFL cohorts were scanned on the Hamamatsu NanoZoomer scanners whilst WSIs from the NKI cohort were scanned on the MIRAX scanners, all at 40x magnification (0·25 µm/pixel). WSIs from the SGK cohort were downloaded from the virtual tissue bank (https://virtualtissuebank.iu.edu/query/), which were scanned at 20x magnification.

To develop *NBT-Classifiers*, we used NKI and BCI cohorts. These slides included samples of NBTs from reduction mammoplasties (n=4 patients, non-*gBRCA1/2m* carriers), risk-reducing mastectomies in *gBRCA1/2m* mutation carriers (n=21 patients, no cancer at surgery), and contralateral or ipsilateral NBTs from breast cancer patients (n=7 patients). The remaining WSIs from the KHP, EPFL and SGK cohorts were used for external validation. These slides included NBTs from reduction mammoplasties (n=26 patients, non-*gBRCA1/2m* carriers) and core biopsies from healthy donors (n=12 patients, non-*gBRCA1/2m* carriers). Detailed information on individual slides can be found in **Supplementary Table 1** and is available at the OASIS repository (https://github.com/cancerbioinformatics/OASIS).

Manual annotation and tessellation

We performed manual annotations using QuPath v0.3.0.[7] To improve the efficiency of the annotation process, we adopted a hybrid strategy: lobular boundaries were precisely delineated to annotate epithelium, while stroma and adipocytes were annotated using rectangular boxes of comparable size, randomly sampled from various regions throughout the NBT, to ensure a balanced representation of tissue classes during training.

We then used QuPath v0.3.0 to extract labelled patches. For the NKI and BCI cohorts, we extracted non-overlapping patches of a size of 224 x 224, 512 x 512 and 1024 x 1024 pixels with a resolution of 0·25 µm/pixel. For cohorts used for external validation (KHP, EPFL and SGK), we only extracted non-overlapping patches with a size of 512 x 512 and 1024 x 1024

pixels with a resolution of 0·25 µm/pixel for the KHP and EPFL cohorts. For the SGK cohort, scanned at 20x magnification, we extracted non-overlapping patches with a size of 256 x 256 and 512 x 512 pixels with a resolution of 0·5 µm/pixel. Only patches with their centres located within the annotated regions were extracted and inherited the labels from the pixel-level manual annotations. The patch extraction was performed using Groovy scripts in QuPath v0.3.0. The script is available at: https://github.com/cancerbioinformatics/NBT-Classifier/blob/main/patching_qupath.groovy.

Architecture of *NBT-Classifiers*

The *NBT-Classifiers* leverage transfer learning, each employing a pre-trained CNN backbone to extract visual features from image patches, followed by a trainable classification head for tissue class prediction (**Supplementary Fig. 4**). Specifically, the feature maps from the last convolutional layer of each CNN backbone were transformed into a single-dimensional vector using a global average pooling layer. The reduced visual features serve as input to the subsequent classification module, which consists of two densely connected layers. The first dense layer linearly transforms the input features into a 1024-dimensional representation, followed by a second dense layer that reduces it to 512 dimensions, both employing the “ReLU” activation function. To mitigate overfitting, a Dropout layer with a rate of 50% is applied between these layers. The final output layer maps the 512-dimensional representation to a 3-dimensional feature, using a “Softmax” activation function to predict the probability distribution across the three tissue classes. The class with the highest probability is assigned as the predicted tissue class (0: epithelium, 1: stroma, 2: adipocytes).

Stain normalisation

In experiments optimising the stain normalisation method, we used the same reference image when implementing the Reinhard, Macenko, and Vahadane method, which is available at: https://github.com/cancerbioinformatics/NBT-Classifier/blob/main/data/he.jpg.

Deep learning implementation

All deep learning experiments were implemented with TensorFlow2 in Python on two NVIDIA A100 GPUs from the high-performance computing cluster of King's Computational Research,

Engineering, and Technology Environment (CREATE). During both training and evaluation, image patches underwent additional pre-processing using input functions tailored to each CNN backbone, provided by the Keras framework. During training, six data augmentation techniques were employed, including horizontal flipping, random rotation within the range of -40 degrees to 40 degrees, random shifts, random shear transformations, random zoom in/out, and random shifts of the channel values within the range of [0, 10], to enhance model's robustness and regularise the algorithm. The weights of the CNN backbones were all frozen, and the densely connected layers in the classification head were initialised by the Xavier normal initializer and set to be trainable. We used categorical cross-entropy as the loss function and Adam optimizer with a learning rate of 1e-5. The batch size was 16 and we trained each model for 50 epochs. The best model with the highest validation accuracy was saved for further testing.

External validation

For cohorts used for external validation (KHP, EPFL and SGK), we extracted non-overlapping patches with a size of 512 x 512 and 1024 x 1024 pixels at a resolution of 0·25 µm/pixel for the KHP and EPFL cohorts. For the SGK cohort, scanned at 20x magnification, we extracted non-overlapping patches with a size of 256 x 256 and 512 x 512 pixels at a resolution of 0·5 µm/pixel. Details of the generated patch-level datasets are listed in **Supplementary Table 2**. All patches were normalised using the Reinhard method and patches from SGK were resized to match the dimension of the corresponding *NBT-Classifier*. For each patch, each *NBT-Classifier* outputs probabilities of epithelium, stroma and adipocytes and the tissue class with the highest predicted probability was assigned as the final class prediction. The primary statistical endpoint was the area under the receiver operating characteristic curve (AUROC). The secondary statistical endpoint was class-specific accuracy. Accuracy was determined by comparing the ground-truth labels with the class assigned the highest predicted probability. We reported both the overall accuracy as well as individual accuracies in each KHP, EPFL and SGK cohort.

Visual assessment of feature representations

We performed t-distributed Stochastic Neighbour Embedding (t-SNE) to visualise feature representations extracted by each *NBT-Classifier* in a 2-dimensional space. To facilitate the inspection of the corresponding histology of the features, we randomly selected 500 H&E

patches from each ground-truth tissue class (epithelium, stroma, adipocytes) and projected them onto the t-SNE plot based on their coordinates. The code for this visualisation is available at: https://github.com/cancerbioinformatics/NBT-Classifier/blob/main/vis_features.ipynb.

Implementation of HistoROI

HistoROI[12] processes 256 x 256-pixel patches at a resolution of 1 µm/pixel, capturing the same field of view (FOV) as 1024 x 1024-pixel patches at 0·25 µm/pixel. To ensure a fair comparison, we resized the 1024 x 1024-pixel patches extracted from WSIs in three external cohorts and applied Reinhard stain normalisation before using them with HistoROI.

Model Interpretability

To enhance model interpretability, we applied Class activation mapping (CAM)[22] and Gradient-weighted Class activation mapping (Grad-CAM).[23] These techniques generate activation heatmaps that highlight the spatial regions most important to the model's predictions for any given input image, utilising the final convolutional layer of the MobileNet backbone, weighted by importance scores. Once the activation heatmaps were computed, we applied bicubic interpolation to match their dimensions to those of the original H&E images. The resulting smoothed heatmaps were overlaid onto the digitised H&E images to visualise the high-attention histological structures corresponding to each particular tissue class prediction. The source code is available at: https://github.com/cancerbioinformatics/NBT-Classifier/blob/main/vis_CAMs.ipynb.

Whole slide tissue classification and visualisation

Given that WSIs often contain substantial non-tissue regions, including white background and artefacts such as coverslip edges, excluding these regions is essential for reducing the model's inference time. Also, as artefacts were not included in the manual annotations and the *NBT-Classifier* tends to misclassify them, this step is crucial to enhance the model's accuracy. We employed HistoQC,[26] a WSI quality control software, to automatically generate binary masks that highlight the artefact-free foreground tissue regions on each WSI, ensuring that only tissue-containing regions are included in downstream tissue classification analyses (**Supplementary Fig. 23**). When using the 512px-based *NBT-Classifier*, the detected tissue region is divided into

non-overlapping patches with a fixed size of 128 x 128 µm to account for variations in the micron per pixel (mpp) values across different scanners. These patches are then resized to 512 x 512 pixels to match the input dimensions of the model. When using the 1024px-based *NBT-Classifier*, the detected foreground region is divided into non-overlapping patches with a fixed size of 256 x 256 µm and resized to 1024 x 1024 pixels. Each *NBT-Classifier* outputs both probabilities of epithelium, stroma and adipocytes and the final class prediction for each patch, along with the original coordination on the slide. To visualise the whole slide tissue classification in a single image, we arranged the class predictions of patches according to their corresponding coordinates on the slide to create a heatmap. For interactive histological inspection, the tissue class heatmap can be exported as a JSON file and imported into the QuPath v0.3.0 platform. Additionally, we generated smoother tissue class heatmaps by upscaling the predicted probability matrix through interpolation. The visualisation method is available at: https://github.com/cancerbioinformatics/NBT-Classifier/blob/main/NBT_pipeline.ipynb.

Lobule detection

To detect individual lobules with their surrounding peri-lobular regions (adjustable), the epithelium layer in the whole slide tissue class heatmap is extracted and enlarged by a factor of 32. Then, connected component analysis is applied to identify groups of connected foreground pixels (filled with a value of 1) as distinct objects. Through manual examination, we observed that lobules empirically are covered by more than two epithelium patches, approximately 500,000 pixels and 0·25 µm/pixel. Therefore, we removed foreground objects (filled with a value of 1) smaller than 400,000 pixels and filled “holes” (filled with a value of 0) smaller than 400,000 pixels inside each detected foreground object. For slides scanned at 20x magnification, the threshold is adjusted to 250,000 pixels. This step reduces noise and increases the robustness for lobule localisation. This method outputs a binary mask highlighting individual lobules and an optional surrounding stroma within a pre-defined range. The mask can be converted into a JSON file and imported into QuPath v0.3.0 for downstream digital image analyses. The source code is available at: https://github.com/cancerbioinformatics/NBT-Classifier/blob/main/NBT_pipeline.ipynb.

Statistical analysis

To compare distributions between subgroups, we performed pairwise Wilcoxon tests, applying Bonferroni correction for multiple comparisons. A p-value of $< 0{\cdot}05$ was considered statistically significant.

Role of the funding source

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

WSIs involved in this study are stored at the OASIS repository: https://github.com/cancerbioinformatics/OASIS, which currently can be accessed upon request.

Code availability

The implementation code involved in this study can be found at: https://github.com/cancerbioinformatics/NBT-Classifier, including Python codes for implementing *NBT-Classifiers* and the pre-processing pipeline, and QuPath v0.3.0 scripts for tessellation and importing pseudo annotations.

Reference

1. Salim M, Liu Y, Sorkhei M, et al. AI-based selection of individuals for supplemental MRI in population-based breast cancer screening: the randomized ScreenTrustMRI trial. *Nature Medicine* 2024; 30: 2623-2630. DOI: 10.1038/s41591-024-03093-5.

2. Henson DE and Tarone RE. Involution and the etiology of breast cancer. *Cancer* 1994; 74: 424-429.

3. Provenzano PP, Eliceiri KW, Campbell JM, et al. Collagen reorganization at the tumor-stromal interface facilitates local invasion. *BMC Med* 2006; 4: 38. 20061226. DOI: 10.1186/1741-7015-4-38.

4. Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *Jama* 2017; 317: 2402-2416.

5. Ferreira R, Moon B, Humphries J, et al. The virtual microscope. In: *Proceedings of the AMIA Annual Fall Symposium* 1997, p.449. American Medical Informatics Association.

6. Pantanowitz L, Sharma A, Carter AB, et al. Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives. *Journal of Pathology Informatics* 2018; 9: 40. DOI: https://doi.org/10.4103/jpi.jpi_69_18.

7. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. *Scientific Reports* 2017; 7: 16878. DOI: 10.1038/s41598-017-17204-5.

8. Hosseini MS, Bejnordi BE, Trinh VQ-H, et al. Computational Pathology: A Survey Review and The Way Forward. *arXiv preprint arXiv:230405482* 2023.

9. Lipkova J and Kather JN. The age of foundation models. *Nat Rev Clin Oncol* 2024; 21: 769-770. DOI: 10.1038/s41571-024-00941-8.

10. Anaya J, Sidhom JW, Mahmood F, et al. Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. *Nat Biomed Eng* 2024; 8: 57-67. 20231102. DOI: 10.1038/s41551-023-01120-3.

11. Zhu Z, Wang SH and Zhang YD. A Survey of Convolutional Neural Network in Breast Cancer. *Comput Model Eng Sci* 2023; 136: 2127-2172. DOI: 10.32604/cmes.2023.025484.

12. Patil A, Diwakar H, Sawant J, et al. Efficient quality control of whole slide pathology images with human-in-the-loop training. *Journal of Pathology Informatics* 2023; 14: 100306.

13. Wetstein SC, Onken AM, Luffman C, et al. Deep learning assessment of breast terminal duct lobular unit involution: Towards automated prediction of breast cancer risk. *PLoS One* 2020; 15: e0231653. 20200415. DOI: 10.1371/journal.pone.0231653.

14. de Bel T, Litjens G, Ogony J, et al. Automated quantification of levels of breast terminal duct lobular (TDLU) involution using deep learning. *NPJ Breast Cancer* 2022; 8: 13. 20220119. DOI: 10.1038/s41523-021-00378-7.

15. Kensler KH, Liu EZF, Wetstein SC, et al. Automated Quantitative Measures of Terminal Duct Lobular Unit Involution and Breast Cancer Risk. *Cancer Epidemiol Biomarkers Prev* 2020; 29: 2358-2368. 20200911. DOI: 10.1158/1055-9965.Epi-20-0723.

16. Ogony J, de Bel T, Radisky DC, et al. Towards defining morphologic parameters of normal parous and nulliparous breast tissues by artificial intelligence. *Breast Cancer Res* 2022; 24: 45. 20220711. DOI: 10.1186/s13058-022-01541-z.

17. Apou G, Schaadt NS, Naegel B, et al. Detection of lobular structures in normal breast tissue. *Comput Biol Med* 2016; 74: 91-102. 20160511. DOI: 10.1016/j.compbiomed.2016.05.004.

18. Abubakar M, Klein A, Fan S, et al. Host, reproductive, and lifestyle factors in relation to quantitative histologic metrics of the normal breast. *Breast Cancer Res* 2023; 25: 97. 20230815. DOI: 10.1186/s13058-023-01692-7.

19. Ish JL, Abubakar M, Fan S, et al. Outdoor air pollution and histologic composition of normal breast tissue. *Environ Int* 2023; 176: 107984. 20230519. DOI: 10.1016/j.envint.2023.107984.

20. Heydarlou H, Hodson LJ, Dorraki M, et al. A Deep Learning Approach for the Classification of Fibroglandular Breast Density in Histology Images of Human Breast Tissue. *Cancers (Basel)* 2025; 17 20250128. DOI: 10.3390/cancers17030449.

21. Tafavvoghi M, Bongo LA, Shvetsov N, et al. Publicly available datasets of breast histopathology H&E whole-slide images: A scoping review. *Journal of Pathology Informatics* 2024; 15: 100363. DOI: https://doi.org/10.1016/j.jpi.2024.100363.

22. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016, pp.2921-2929.

23. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision* 2020; 128: 336-359.

24. Sherman ME, Figueroa JD, Henry JE, et al. The Susan G. Komen for the Cure Tissue Bank at the IU Simon Cancer Center: A Unique Resource for Defining the “Molecular Histology”

of the Breast. *Cancer Prevention Research* 2012; 5: 528-535. DOI: 10.1158/1940-6207.CAPR-11-0234.

25. Hoque MZ, Keskinarkaus A, Nyberg P, et al. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion* 2024; 102: 101997. DOI: https://doi.org/10.1016/j.inffus.2023.101997.

26. Janowczyk A, Zuo R, Gilmore H, et al. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clin Cancer Inform* 2019; 3: 1-7. DOI: 10.1200/cci.18.00157.

27. Nejat P, Alsaafin A, Alabtah G, et al. Creating an atlas of normal tissue for pruning WSI patching through anomaly detection. *Scientific Reports* 2024; 14: 3932. DOI: 10.1038/s41598-024-54489-9.

28. Montezuma D, Oliveira SP, Tolkach Y, et al. Annotation Practices in Computational Pathology: A European Society of Digital and Integrative Pathology (ESDIP) Survey Study. *Lab Invest* 2024; 105: 102203. 20241129. DOI: 10.1016/j.labinv.2024.102203.

29. Reed AD, Pensa S, Steif A, et al. A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nature Genetics* 2024; 56: 652-662. DOI: 10.1038/s41588-024-01688-9.

Acknowledgements

The authors wish to acknowledge the support of the King's Health Partners Cancer Biobank in London, the Netherlands Cancer Institute in Amsterdam, the Barts Cancer Institute in London, and the École Polytechnique Fédérale de Lausanne in Lausanne, Switzerland and Aasiyah Oozeer, Marcello D'Angelo, Rachel Barrow, Rachel Nelan, Marcelo Sobral-Leite, Fabio de Martino for material collection. The authors would like to thank all members of the Cancer Bioinformatics team at King’s College London (London, UK) for their helpful suggestions. We thank the Breast Cancer Research Trust, Breast Cancer Now (and their legacy charity Breakthrough Breast Cancer), the Medical Research Council (MRC) [MR/X012476/1], Cancer Research UK [CRUK/07/012, KCL-BCN-Q3], and CRUK City of London Centre Award [CTRQQR-2021/100004] for funding this project. Siyuan Chen is funded by a China Scholarship Council PhD scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. During the preparation of this work the author(s) used Grammarly (free version) and ChatGPT (version 2) in order to correct grammatical errors and enhancing the overall readability of the manuscript. After using this

tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Author Contributions

Conceptualisation: Anita Grigoriadis, Sarah Pinder, Siyuan Chen, Mario Parreno-Centeno, Christopher R.S. Banerji. **Data curation**: Siyuan Chen, Graham Booker. **Formal analysis**: Siyuan Chen, Mario Parreno-Centeno. **Funding acquisition**: Anita Grigoriadis. **Investigation**: Anita Grigoriadis, Sarah Pinder, Siyuan Chen, Mario Parreno-Centeno, Christopher R.S. Banerji. **Methodology**: Siyuan Chen, Mario Parreno-Centeno, Christopher R.S. Banerji, Gregory Verghese, Salim Arslan. **Project administration**: Anita Grigoriadis. **Resources**: Cheryl Gillett, Louise J. Jones, Esther Lips, Cathrin Brisken, Aasiyah Oozeer, Marcello D'Angelo, Rachel Barrow, Rachel Nelan, Marcelo Sobral-Leite, Fabio de Martino. **Software**: Siyuan Chen, Mario Parreno-Centeno. **Supervision**: Anita Grigoriadis, Sarah Pinder, Christopher R.S. Banerji. **Validation**: Fathima Sumayya Mohamed. **Visualisation**: Siyuan Chen. **Writing – original draft**: Siyuan Chen. **Writing– review & editing**: Anita Grigoriadis, Sarah Pinder, Christopher R.S. Banerji, Esther Lips, Cathrin Brisken, Louise J. Jones, Mathew J. Smalley, Gregory Verghese, Fathima Sumayya Mohamed, Graham Booker, Siyuan Chen, Mario Parreno-Centeno, Salim Arslan, Pahini Pandya.

Competing interests

Anita Grigoriadis, Louise J. Jones and Greg Verghese are Co-Founders of PharosAI. Salim Arslan and Pahini Pandya and employed by Panakeia Technology, UK. All other authors declare no conflict of interest.

Figures

Figure 1

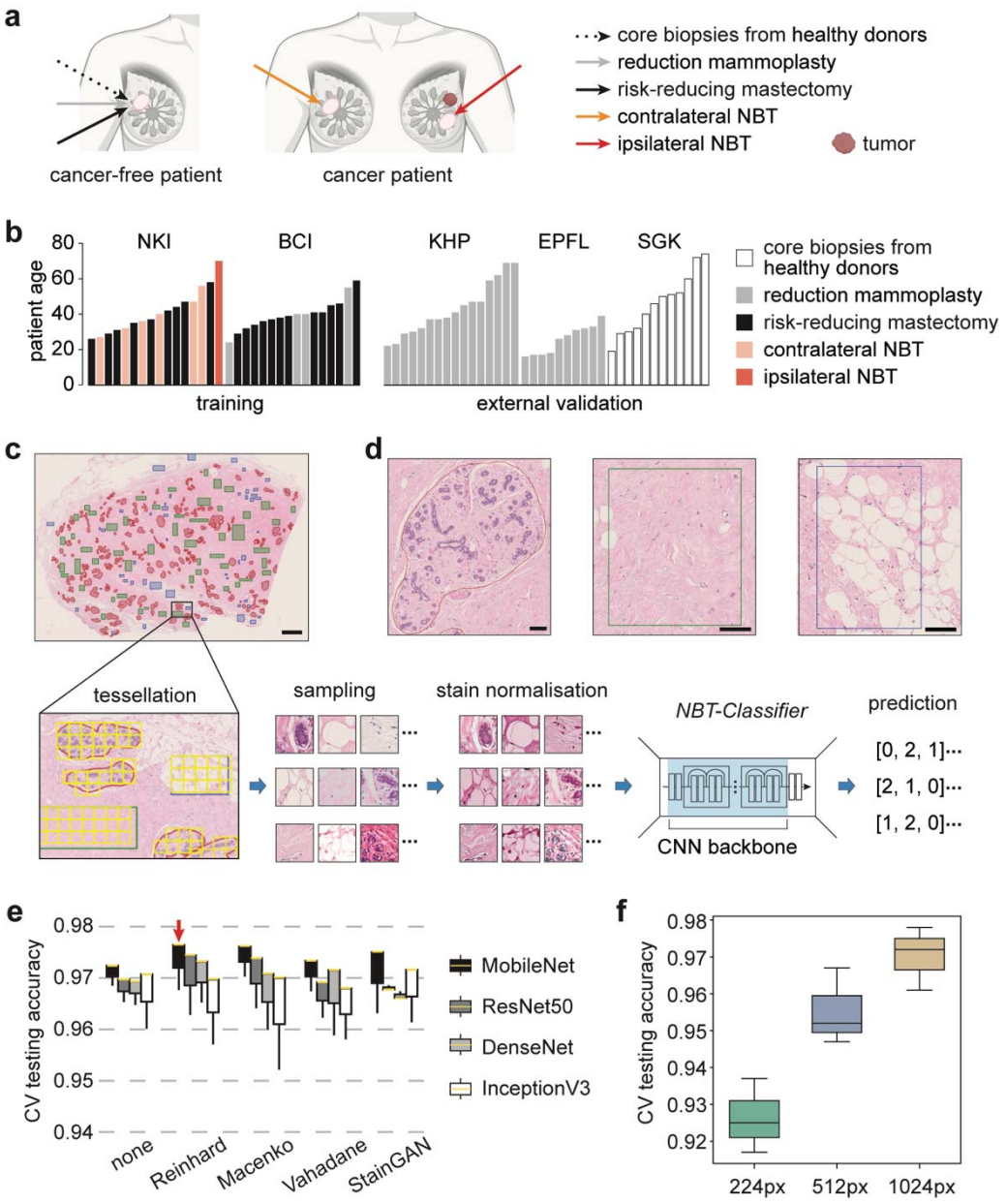

Figure 1. Dataset overview and optimisation of *NBT-classifiers.* **a** Schematic illustrating the sources of normal breast tissues (NBTs) used in the study. **b** Barplot depicting the distribution of whole slide images (WSIs) used for training and external validation. WSIs from the NKI and BCI cohorts were used for model training, while those from the KHP, EPFL, and SGK cohorts were reserved for external validation. Each bar represents the patient age associated with a single WSI, with colours indicating the NBT source. **c** Schematic of the manual annotation process and the training pipeline. Epithelium, stroma,

and adipocytes are annotated in red, green, and blue, respectively. Scale bar, 2 mm. Annotations were tessellated into non-overlapping patches, followed by class-balanced sampling and stain normalisation for training each *NBT-Classifier*. Each model uses a CNN backbone with a classification head that predicts three tissue classes: 0 (epithelium), 1 (stroma), and 2 (adipocytes). **d** Magnified images of single annotations for epithelium, stroma, and adipocytes (from left to right). Scale bars, 100 µm. **e** Barplot of 3-fold cross-validation (CV) performance for models trained with and without different stain normalisation methods and CNN backbones. Experiments were conducted on patches tessellated at 512 x 512 pixels (0·25 µm/pixel). The yellow line in each box represents the median internal test accuracy, with whiskers extending to the minimum and maximum values. The best-performing combination is highlighted with a red arrow. **f** Barplot showing test accuracies during 3-fold cross-validation for models trained on different patch sizes. Patches were tessellated at 0.25 µm/pixel resolution, processed using the Reinhard stain normalisation method and the MobileNet backbone. The results for 224 x 224 (green, 224px), 512 x 512 (blue, 512px), and 1024 x 1024 (light brown, 1024px) pixel patches are coloured differently. The upper, middle, and lower lines of each boxplot represent the test accuracies during 3-fold cross-validation.

Figure 2

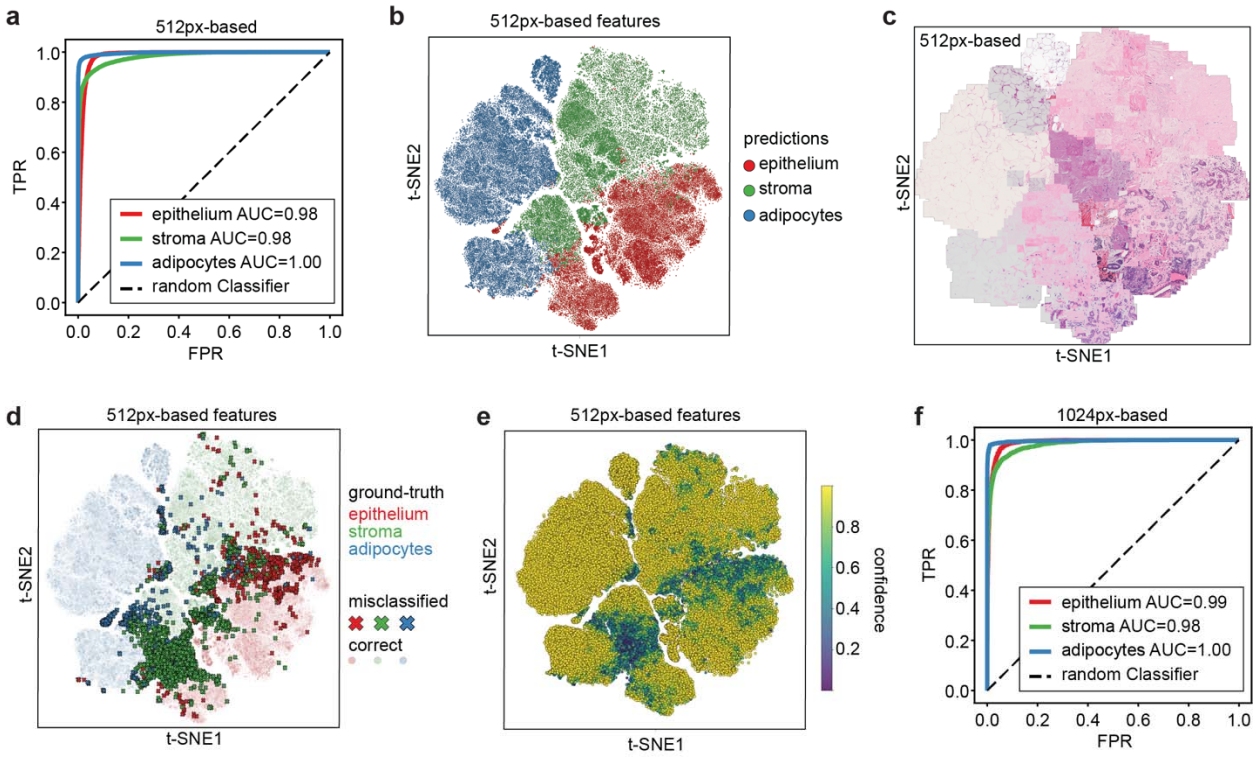

Figure 2. External validation performance. a Multi-class ROC curves showing the external performance of the 512px-based *NBT-Classifier*. True positive rate (TPR) is plotted against false positive rate (FPR) at different thresholds for epithelium, stroma, and adipocytes, indicated in red, green, and blue, respectively. The corresponding area under the curve (AUC) values are presented in the legend. **b** t-SNE visualisation of the 512px-based *NBT-Classifier's* predictions. Features for epithelium, stroma, and adipocytes are indicated in red, green, and blue, respectively. **c** t-SNE visualisation of H&E-stained patches of the same samples shown in panel b. For each predicted tissue class (epithelium, stroma, and adipocytes), 500 patches were randomly sampled and overlaid onto the plot based on their coordinates. **d** t-SNE visualisation of misclassified patches by the 512px-based *NBT-Classifier*. The plot is coloured by ground-truth tissue classes, with misclassified patches highlighted as opaque crosses (alpha=1.0) and correctly classified patches shown as semi-transparent circles (alpha=0.3). **e** t-SNE visualisation of the prediction confidence of the 512px-based *NBT-Classifier*. Each patch is coloured based on the probability associated with its predicted tissue class. **f** Multi-class ROC curves showing the external performance of the 1024px-based *NBT-Classifier*.

Figure 3

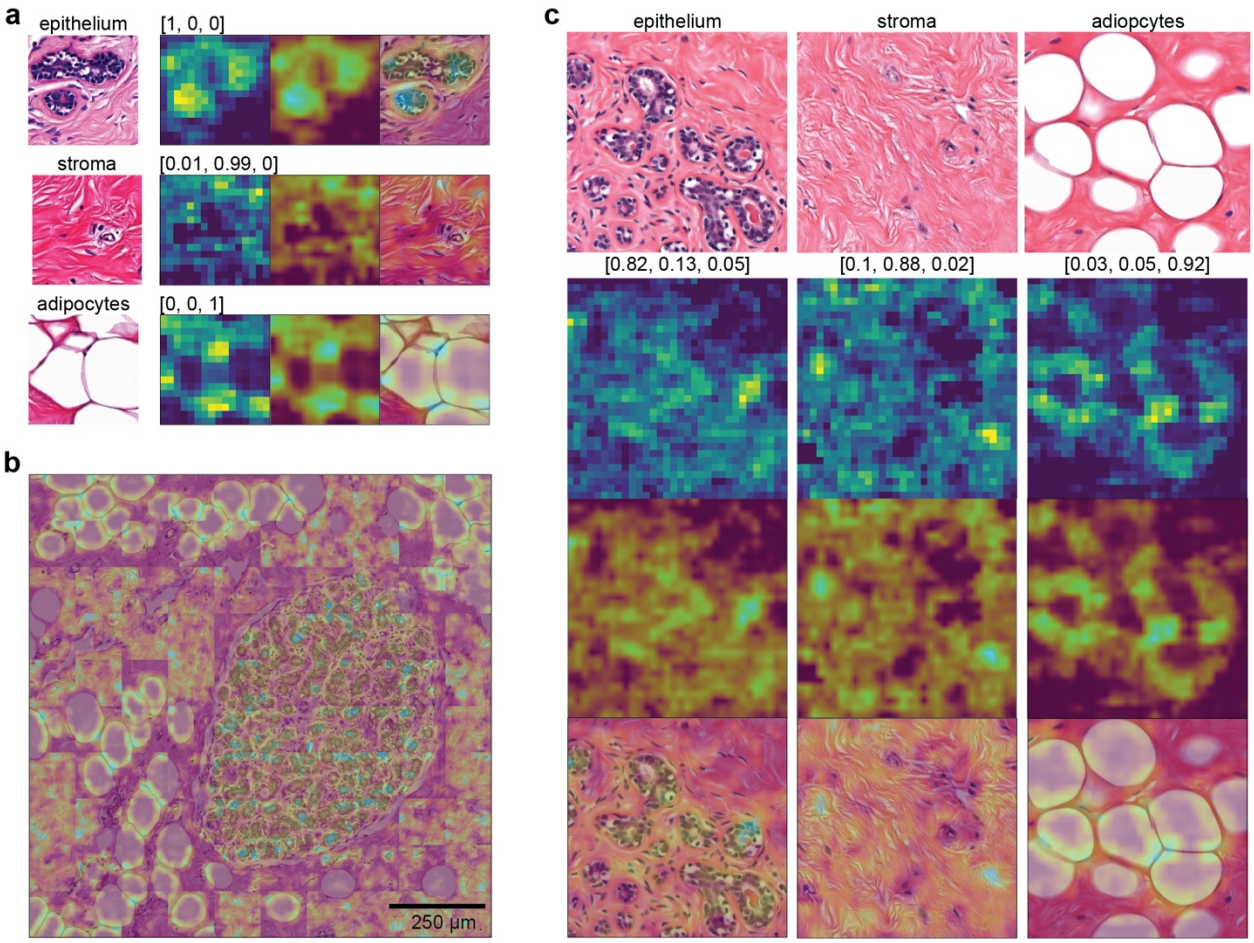

Figure 3. Interpretation of *NBT-Classifiers'* predictions. a Class activation mapping (CAM) visualisation for the 512px-based *NBT-Classifier*. Examples of epithelium, stroma, and adipocytes patches (512 x 512 pixels, 0·25 µm/pixel) are shown in the first column. For each example, the original CAM heatmap, resized heatmap and heatmap overlaid on the original H&E patch image are shown on the right, with the predicted probabilities for epithelium, stroma, and adipocytes displayed at the top. **b** CAM heatmaps overlaid on a selected region. The region was divided into non-overlapping 512 x 512-pixel patches, for which CAM heatmaps were generated based on the predicted tissue class. The heatmaps were then resized and overlaid to highlight high-attention structures in each patch. Scale bar, 250 µm. **c** CAM visualisation for the 1024px-based *NBT-Classifier*. Examples of epithelium, stroma, and adipocytes patches (1024 x 1024 pixels, 0·25 µm/pixel) are shown in the first row, with the predicted probabilities of epithelium, stroma, and adipocytes beneath each image. For each example, the original CAM heatmap, resized heatmap and heatmap overlaid on the original patch are displayed beneath.

Figure 4

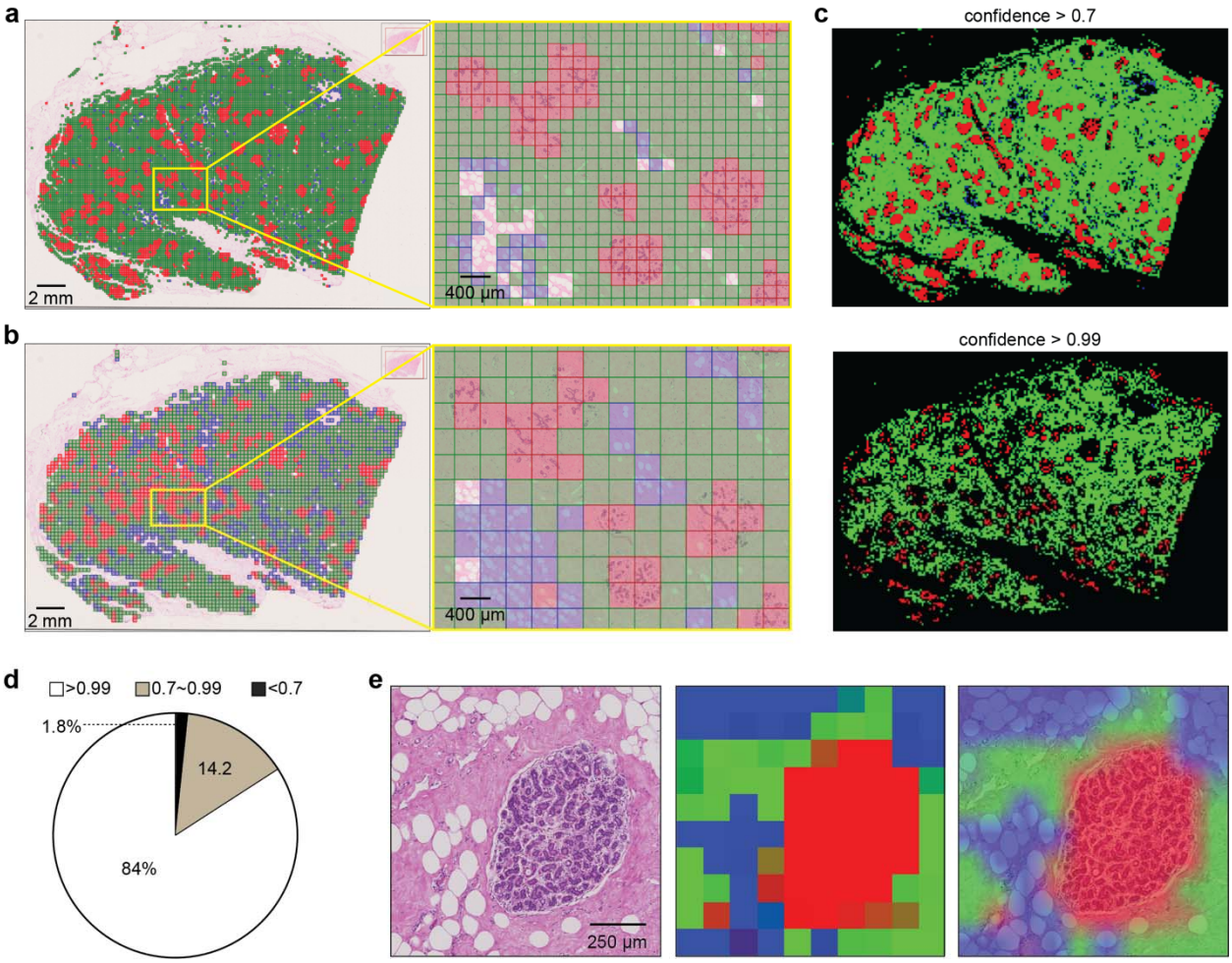

Figure 4. Visualisation of whole slide tissue classification. a-b QuPath v0.3.0 visualisation of whole slide tissue class predictions by the 512px- (**a**) and 1024-based (**b**) *NBT-Classifiers*. The original tissue class predictions of the entire WSI were converted into JSON files and imported into QuPath v0.3.0 as pseudo patch-level annotations, shown in the right images. Epithelium, stroma, and adipocytes are represented in red, green, and blue, respectively. Scale bar, 2 mm. Magnified regions are displayed on the right. Scale bar, 400 µm. **c** Visualisation of patches predicted with varying confidence. The top image shows patches classified with a confidence higher than 0.7, while the bottom image shows patches classified with a confidence higher than 0.99. **d** Pie chart illustrating the distribution of the 512px-based *NBT-Classifier's* confidence in the predicted tissue class. **e** Tissue probability heatmaps of a representative region. A 5120 x 5120-pixel region containing mixed tissue components (left) was selected and tessellated into non-overlapping 512 x 512-pixel patches. For each patch, the probabilities of epithelium, stroma, and adipocytes were predicted using the 512px-based *NBT-Classifier* and mapped to

the red (R), green (G), and blue (B) channels, respectively, to generate a tissue probability heatmap (middle). The heatmap was then resized and overlaid onto the H&E image to facilitate comparison with the corresponding histological patterns (right). Scale bar, 250 µm.

Figure 5

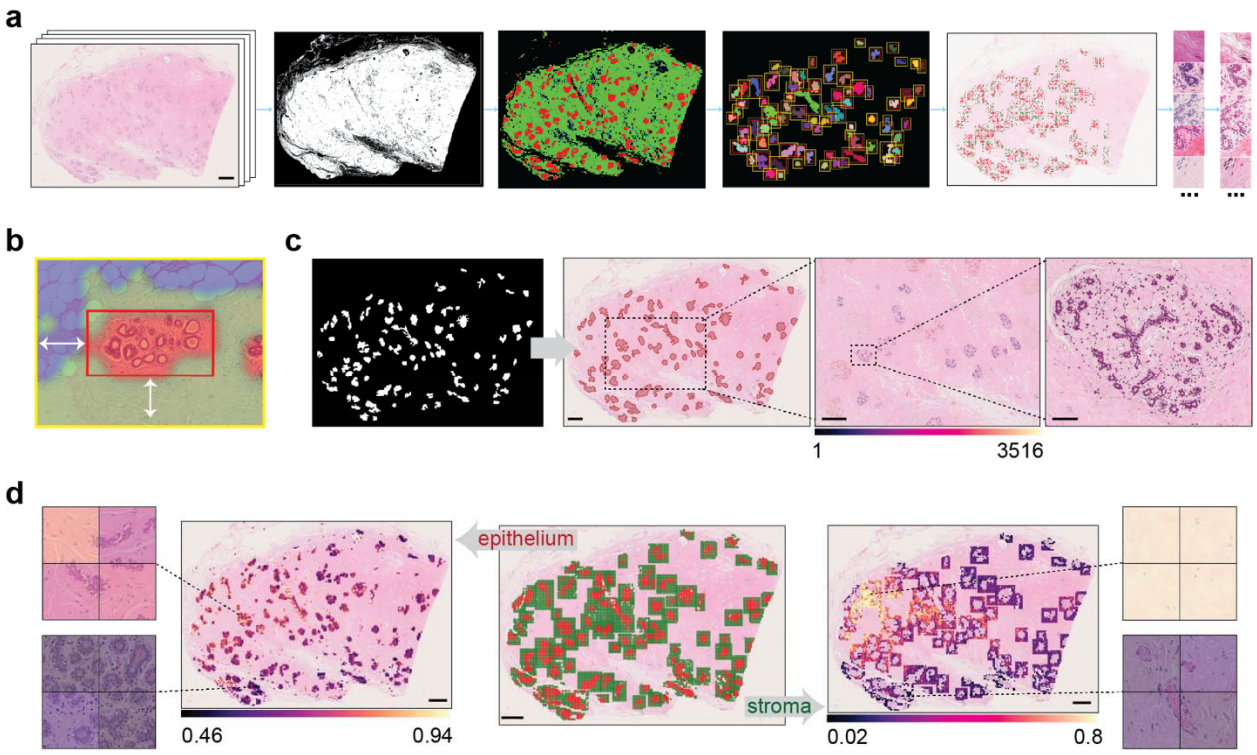

Figure 5. Approaches facilitating downstream WSI analysis of NBTs. a Proposed WSI pre-processing pipeline for NBTs. From left to right, the images depict an input WSI (scale bar, 2 mm), detected foreground tissue regions, whole slide tissue classifications (red: epithelium, green: stroma, blue: adipocytes), identified lobular regions, balanced-sampled patches, and stain-normalised patches. **b** Magnified example of a detected lobular region. The tissue class mask was generated by the 512px-based *NBT-Classifier*, resized, and overlaid onto the H&E image. The red box highlights the detected lobule, while the yellow box outlines the lobular region, with a 250 µm (adjustable) width between the two boxes, as indicated by the white arrows. **c** Example demonstrating generated pseudo lobule annotations and downstream nuclei spatial analysis in QuPath v0.3.0. The first image displays a binary mask of detected lobules, which were imported as pseudo-lobule annotations (highlighted in red) in QuPath v0.3.0, as shown in the second image. Scale bar, 2 mm. Within QuPath v0.3.0, nuclei segmentation was performed using StarDist, followed by Delaunay spatial analysis to quantify various graph properties of each lobule, such as the number of nuclei within the graph. The third image visualises the variation in the number of nuclei (indicated by the colour bar) across Delaunay graphs constructed for lobules within a magnified region. Scale bar, 1 mm. The last image presents an example of a Delaunay graph constructed from a magnified lobule. Scale bar, 100 µm. **d** Example of pseudo patch-level tissue class annotations within detected lobular regions and downstream texture analysis in QuPath v0.3.0. The middle image displays pseudo patch-level annotations of the epithelium (red) and peri-lobular stroma (green) in QuPath

v0.3.0, which were subsequently analysed separately using Haralick texture features. The left image visualises the variation in the Inverse Difference Moment (F4) across epithelium patches, while the right image shows the variation in the Angular Second Moment (F0) across stroma patches. Scale bar, 2 mm. Magnified patch examples with low and high F4 and F0 values are displayed on the side.

Supplementary Information

Supplementary Figure 1: Schematic pipeline of the manual annotation procedure using QuPath v0.3.0. For each WSI, epithelium regions (in red) were exhaustively delineated using the brush tool to ensure precise annotation. To obtain a balanced number of training patches from each tissue class, similarly sized regions of stroma (in green) and adipocytes (in blue) were randomly selected from various areas of the WSI using rectangular boxes. This approach enhances efficiency while ensuring the inclusion of spatial variations and maintaining proportional representation across tissue types.

Supplementary Figure 2: Illustration of the histology of NBTs captured at different fields of view. From left to right, the columns display example patches of epithelium, stroma, and adipocytes. Within each column, the rows, from top to bottom, show histological structures captured at fields of view of 256 x 256 µm, 128 x 128 µm, and 64 x 64 µm, respectively.

Supplementary Fig. 3 Examples of annotated WSIs of NBTs from BCI, EPFL, NKI, SGK and KHP cohorts. Epithelium, stroma, and adipocytes are annotated in red, green, and blue, respectively. Scale bar, 2 mm.

Supplementary Figure 4: Illustration of *NBT-Classifiers'* architecture. Each *NBT-Classifier* utilizes a pre-trained CNN backbone to extract feature maps, which are then processed via global average pooling. A classification head subsequently predicts the probabilities for epithelium, stroma, and adipocytes, with the final tissue class assigned based on the highest probability: 0 for epithelium, 1 for stroma, and 2 for adipocytes.

Supplementary Figure 5: Illustration of different stain normalisation methods*.* The first column shows the original H&E patch images, highlighting considerable staining variations across KHP, NKI, BCI, EPFL and SGK cohorts. Columns two through five display the patches after normalisation using the Reinhard, Vahadane, Macenko, and StainGAN methods, demonstrating varying abilities to mitigate staining variations.

Supplementary Figure 6: Barplot showing the number of patches obtained using different patch sizes. The panels illustrate the variation in the number of patches obtained from the NKI (top) and BCI (bottom) cohorts, using patch sizes of 224 x 224 pixels (denoted as 224px) (left), 512 x 512 pixels (denoted as 512px) (middle), and 1024 x 1024 pixels (denoted as 1024px) (right) at 40x magnification. In each panel, the x-axis represents individual WSI, with the coloured bar plots separately showing the number of epithelium (red), stroma (green), and adipocytes (blue) patches generated from each WSI at the corresponding patch size. The horizontal line indicates the number of 250 patches, while the horizontal dashed line indicates the number of 150 patches.

Supplementary Figure 7: Sample workflow for three-fold cross-validation experiments. A total of 23 training configurations were evaluated, varying in stain normalisation methods, CNN backbones, and patch sizes.

Supplementary Figure 8: Confusion matrices evaluating the 224px-, 512px and 1024px-based *NBT-Classifiers* on external datasets. Each confusion matrix summarizes the classification results across the three external cohorts (KHP, EPFL, and SGK), comparing the predicted tissue classes: epithelium, stroma, and adipocytes (columns) against the ground-truth labels (rows). Diagonal cells indicate the number of correct predictions for each class, while off-diagonal cells indicate misclassifications. The colour intensity of each cell reflects the classification frequency, with lighter shades indicating higher counts. The overall balanced accuracy is derived from these matrixes and displayed at the top.

Supplementary Figure 9: Cohort-specific multi-class receiver operating characteristic curves for the 512px-based *NBT-Classifier*. Panels from left to right show the receiver operating characteristic (ROC) curves of the 512px-based NBT-Classifier evaluated in the KHP, EPFL and SGK cohort, respectively. Performance of epithelium, stroma and adipocytes are indicated in red, green, and blue, respectively, with corresponding values of area under the ROC curve (AUC) presented below each curve.

Supplementary Figure 10: t-SNE visualisation of features extracted by the 512px-based *NBT-Classifier*, coloured by cohort. Each point represents a patch from three external datasets: KHP, EPFL, and SGK. The features are derived from the last layer (the layer before the output layer) of the 512px-based *NBT-Classifier*.

Supplementary Figure 11: Multi-class ROC curves for HistoROI. ROC curves for epithelium, stroma and adipocytes are indicated in red, green and blue, respectively. The corresponding AUCs are presented at the bottom.

Supplementary Figure 12: Cohort-specific multi-class ROC curves for the 1024px-based *NBT-Classifier*. Panels from left to right show the ROC curves of the 1024px-based *NBT-Classifier* evaluated in the KHP, EPFL and SGK cohort, respectively. Performance of epithelium, stroma and adipocytes are indicated in red, green, and blue, respectively, with corresponding values of AUC presented below each curve.

Supplementary Figure 13: Cohort-specific multi-class ROC curves for HistoROI. Panels from left to right show the ROC curves of the HistoROI model evaluated in the KHP, EPFL and SGK cohorts, respectively. Performance of epithelium, stroma and adipocytes are indicated in red, green, and blue, respectively, with corresponding values of AUC presented below each curve.

Supplementary Figure 14: Proportional confusion matrices for the 1024px-based *NBT-Classifier* and HistoROI. For HistoROI, representative examples of epithelium and misclassified as artefact are indicated in red, stroma and misclassified as artefact are indicated in yellow, adipocytes and misclassified as artefact are indicated in brown, and adipocytes and misclassified as miscellaneous are indicated in green.

Supplementary Figure 15: Visualisation of features extracted by the 1024px-based *NBT-Classifier*. The top-left panel displays a t-Distributed Stochastic Neighbour Embedding (t-SNE) plot coloured by the 1024px-based *NBT-Classifier's* predictions. The top-right panel displays a t-SNE plot with points coloured according to ground-truth tissue classes. Misclassified patches are marked as opaque crosses (alpha=1.0), while correctly classified patches are represented by semi-transparent circles (alpha=0.3). The bottom-left panel overlays H&E-stained patches onto a subset of random patches on the t-SNE plot. The bottom-right panel shows a t-SNE plot coloured by cohort, illustrating domain shift across cohorts.

Supplementary Figure 16: Visualisation of features extracted by HistoROI. The top-left panel displays a t-SNE plot coloured by the HistoROI model's predictions. The top-right panel displays a t-SNE plot with points coloured according to ground-truth tissue classes. Misclassified patches are marked as opaque crosses (alpha=1.0), while correctly classified patches are represented by semi-transparent circles (alpha=0.3). The bottom-left panel overlays H&E-stained patches onto a subset of random patches on the t-SNE plot. The bottom-right panel shows a t-SNE plot coloured by cohort, illustrating domain shift across cohorts.

Supplementary Figure 17: Boxplot comparisons of *NBT-Classifiers'* performance across different age groups and NBT sources. The performance of the 512px- and 1024px-based models is depicted in white and black boxplots, respectively. In each boxplot, the central line indicates the median, while the whiskers extend to the minimum and maximum values. The label "ns" denotes statistical non-significance, with adjusted p-values less than 0·05 considered significant. RM stands for reduction mammoplasty, and RRM stands for risk-reducing mastectomy.

Supplementary Figure 18: Illustration of class activation mapping and gradient-weighted class activation mapping visualisations. Class activation mapping (CAM) (Panel **a**) uses class-specific weights to aggregate feature maps from the last convolutional layer of the CNN backbone, while gradient-weighted class activation mapping (Grad-CAM) (Panel **b**) computes the average gradients. Both methods highlight high-attention regions that are relevant to a specific tissue class.

Supplementary Figure 19: Grad-CAM visualisations. Panel **a** displays Grad-CAM heatmaps for 512 x 512-pixel patches (0·25 µm/pixel) of the epithelium (top row), stroma (middle row), and adipocytes (bottom row). Panel b shows visualisations for 1024 x 1024-pixel patches (0·25 µm/pixel) of the same

tissue types. Each row, from left to right, includes the original Grad-CAM heatmaps, resized heatmaps, and heatmaps overlaid on the original patches.

Supplementary Figure 20: Consistent histopathological patterns are captured across multiple cohorts. The rows indicate CAM heatmaps overlaid on the Reinhard normalised epithelium, stroma and adipocytes patches (512 x 512 pixels, 0·25 µm/pixel) from KHP, BCI and EPFL cohorts. Each patch was predicted by the 512px-based *NBT-Classifier* with a probability of 1·0 of the predicted tissue class. The highlighted histopathological patterns in each tissue class are consistent across multiple cohorts.

Supplementary Figure 21: CAM and Grad-CAM heatmaps for epithelium, stroma and adipocytes patches from the SGK cohort. These patches are 256 x 256 pixels, 0·5 µm/pixel, obtained from WSIs in the SGK cohort that were scanned at 20x magnification.

Supplementary Figure 22: CAM-based interpretation of low-confidence predictions by the 512px-based *NBT-Classifier*. The rows, from top to bottom, show low-confidence 512 x 512-pixel patches predicted as epithelium, stroma, and adipocytes, respectively. For each patch, CAM heatmaps are generated for all three tissue classes and overlaid on the corresponding H&E patch image, highlighting class-specific high-attention regions. The predicted class probabilities are displayed at the top, with epithelium in red, stroma in green, and adipocytes in blue.

Supplementary Figure 23: Illustration of foreground tissue detection using HistoQC. The top left image shows the original H&E-stained WSI of NBTs. The top right image shows the foreground tissue (in white) detected by the conventional Otsu threshold algorithm. Notably, the coverslip edge is incorrectly included as part of the foreground tissue. The bottom left image shows the foreground tissue (in white) detected by HistoQC, with the coverslip edge (in white) separately identified in the bottom right image.

Supplementary Figure 24: Representative regions with overlaid tissue probability heatmaps. Heatmaps were generated from NBTs with distinct patient ages and NBT sources, as indicated at the top of each figure. RRM: risk-reducing mastectomy; RM: reduction mammoplasty. Each heatmap delineates the epithelium (red), stroma (green), and adipocytes (blue). Scale bar, 250 µm.

Supplementary Figure 25: Examples illustrating the proposed *NBT-Classifier*-based WSI pre-processing pipeline. The top row shows a WSI example from the NKI cohort, scanned at 40x magnification, while the bottom row shows a WSI example from the SGK cohort, scanned at 20x magnification. From left to right, the images display the original WSI, the whole slide tissue classification heatmap, detected lobular regions (lobules are localised by the inner red boxes and peri-lobular regions are localised by the outer yellow boxes), and pseudo-patch-level annotations (red for epithelium patches and green for stroma patches) visualised in QuPath v0.3.0.

Supplementary Table 1. Overview of annotated WSI datasets of NBTs

Supplementary Table 2. Summary of annotated patch-level datasets

Supplementary Table 3. A summary of studies of NBTs with manual annotations in recent ten years