Predicting Endometrial Cancer Subtypes and Molecular Features from Histopathology Images Using Multi-resolution Deep Learning Models

Runyu Hong[1,2], Wenke Liu[1,2], Deborah DeLair[3], Narges Razavian[4], David Fenyö[1,2]
David.Fenyo@nyulangone.org
1.Institute for Systems Genetics, NYU School of Medicine, New York, NY 10016, USA
2.Department of Biochemistry and Molecular Pharmacology, NYU School of Medicine, New York, NY 10016, USA
3.Department of Pathology, NYU Langone Health, New York, NY 10016, USA
4.Department of Population Health, NYU Langone Health, New York, NY 10016, USA

Abstract

The determination of endometrial carcinoma histological subtypes is a critical diagnostic process that directly affects patients' prognosis and treatment options. Recently, molecular subtyping and mutation status are increasingly utilized in clinical practice as they offer better inform prognosis and offer the possibility of individualized therapies. Compared to the histopathological approach, however, the availability of molecular subtyping is limited as it can only be obtained by genomic sequencing, which may be cost prohibitive. Here, we implemented deep convolutional neural network models that predict not only the histological subtypes, but also molecular subtypes and 18 common gene mutations based on digitized H&E stained pathological images. Taking advantage of the multi-resolution nature of the whole slide images, we introduced a customized architecture, Panoptes, to integrate features of different magnification. The model was trained and evaluated with images from The Cancer Genome Atlas and Clinical Proteomic Tumor Analysis Consortium. Our models achieved an area under the receiver operating characteristic curve (AUROC) of 0.969 in predicting histological subtype and 0.934 to 0.958 in predicting the copy number high (CNV-H) molecular subtype. The prediction tasks of 4 mutations and microsatellite high (MSI-H) molecular subtype also achieved a high performance with AUROC ranging from 0.781 to 0.873. Panoptes showed a significantly better performance than InceptionResnet in most of these top predicted tasks by up to 18%. Feature extraction and visualization revealed that the model relied on human-interpretable patterns. Our results suggest that Panoptes can help pathologists determine molecular subtypes and mutations without sequencing, and our models are generalizable to independent datasets.

Introduction

Endometrial cancer is the most common type of gynecologic cancer among women around the world with a rising occurrence and mortality[1–4]. In the United States, it is one of the top 5 leading cancer types with 52,600 new cases reported in 2014. This number increased to 60,050 in the year of 2016, and was estimated to further increase to 61,880 in 2019[3–6]. Globally, endometrial cancer caused approximately 42,000 women's death in 2005, and this annual mortality count estimate drastically increased to 76,000 in 2016[1,2]. The 5-year survival rate, depending on the study cohort, is ranging from 74% to 91% for patients without metastasis[3].

Clinically, endometrial carcinomas are stratified based on their grade, stage, hormone receptor expression, and histological characteristics[7]. Histological classification reflects tumor cell type and informs the choice of surgical procedure and adjuvant therapy. The majority of endometrial cancer cases exhibit either endometrioid or serous characteristics, which comprise approximately 70%-80% and about 10% of all cases, respectively[8]. Statistically, patients with serous subtype tumors have a lower 5-year survival rate due to more frequent metastases and a higher risk of recurrence[2]. Thus, it is critical to determine the subtypes in order to determine patients' individualized treatment plans and to assess prognosis[1,9]. Histological subtype is determined by pathologists after thorough examination of hematoxylin and eosin (H&E) stained tissue sample slides of tumor samples Endometrioid tumors typically exhibit a glandular growth pattern, while the serous subtype is characterized by the frequent presence of a complex papillary pattern[10–12]. These features are not exclusive for either of the subtypes, however, making histological classification challenging, especially among high grade cases, even for experienced pathologists and necessitating ancillary subtyping criteria[2,13–15].

The multi-omics study of The Cancer Genome Atlas (TCGA) introduces a set of novel criteria that classify endometrial carcinoma into four molecular subtypes, namely POLE ultra-mutated, high microsatellite instability (MSI-H) hypermutated, copy-number low (CNV-L), and copy-number high (CNV-H), based on their mutation characteristics, copy number alterations, and microsatellite instability. This molecular classification standard has been gaining popularity among pathologists and clinicians in recent years. Among these four subtypes, patients with the CNV-H subtype, which includes serous carcinomas and a subset of high grade endometrioid cancers, had the worst outcomes based on progression free survival[15]. Exome sequencing also revealed a panel of genes differentially mutated across the four molecular subtypes, many of which have been shown to play significant roles in endometrial carcinoma tumorigenesis and proliferation and can potentially be novel targets of individualized therapies[16,17]. For example, most patients in CNV-H subtype are TP53 mutated but PTEN wild-type. Determining the molecular subtyping and single gene mutations can provide new insights that complement and refine the histological classification, but the availability of this information is limited by the time and cost of sequencing.

New powerful computational approaches for analyzing massive biomedical data have tackled numerous challenges, which accelerates the pace of human health improvement worldwide. In particular, computational pathology, a discipline that involves the application of image processing techniques to pathological data, has been especially benefitted from the advancement of deep learning in recent years[18–21]. Convolutional neural network (CNN) models are capable of segmenting cells in histopathology slides and classifying them into different types based on their morphology[18,22]. An InceptionV3-based model achieves a high level of accuracy in determining melanoma possibility, exhibiting significant diagnostic potential[19]. Moreover, successful deep learning models have also been built to predict molecular and genomic features in cancer, such as microsatellite instability (MSI) and somatic mutation status, suggesting that machine learning techniques may be able to assist human experts to further exploit clinically relevant information in pathological images[23,24].

H&E slides are often scanned at multiple resolutions (20X, 10X, 5X) and different resolutions of the same slides are saved into a single image file. This allows pathologists to examine features of various sizes at the optimal resolution. Here, we designed a customized architecture, that we call Panoptes. Panoptes takes advantage of the multi-resolution structure of the H&E image files. We showed that models using this architecture could classify endometrial carcinoma histological subtypes, CNV-H and MSI-H molecular subtypes, and several critical mutations with decent performance based on H&E images and outcompete existing InceptionResnet models in most top-performing tasks. Using tSNE dimensionality reduction techniques, we extracted and visualized the features learned by models to classify H&E images. These histopathological features were mostly human interpretable, suggesting possibilities of incorporating them into the pathological diagnostic standards. In particular, we confirmed that tumor grade was the major factor to distinguish CNV-H molecular subtype from the other 3 molecular subtypes in the histological endometrioid cases.

Results

Data preparation and multi-resolution deep-learning based histopathology image analysis. The goal of this study was to build multi-resolution deep convolutional neural network models that could automatically analyze endometrial cancer digital H&E slides and predict their histological and molecular features. We used diagnostic formalin-fixed paraffin-embedded (FFPE) and H&E stained tumor slides and labels from 2 public databases, the GDC data portal containing data of TCGA, and The Cancer Imaging Archive (TCIA) containing data of Clinical Proteomic Tumor Analysis Consortium (CPTAC), to train and test our models. TCGA and CPTAC are two mutually independent cohorts. 107 slides from 98 patients belonged to the CPTAC cohort and 389 slides from 358 patients were in the TCGA cohort. Overall, 496 slides from 456 patients, covered in previous publications (Dou et al., 2020; Getz et al., 2013) and annotated with subtype and gene mutation information, were included to form a mixed dataset (Fig. 1a, Supplementary Fig. 1a). More than 90% of patients in our cohort had only 1 diagnostic slide (Supplementary Fig. 1b). As a lot of driver gene mutations in endometrial cancer are correlated with histological and molecular subtypes, we validated these correlations to ensure that our cohort was a representative of the patient's population (Supplementary Fig. 1c). The general process of training, validation, testing, and visualization followed the workflow in Fig. 1b. For each prediction task, cases in the mixed dataset were randomly split into training, validation, and test set such that slides from the same patient were in only one of these sets. This allowed the test set to be strictly independent of the training process and also made it possible to obtain per-patient level metrics, which could be more useful in the clinical setting. Each task was performed on a different random split of cases stratified with the outcome. Due to the extremely large dimension of the digital H&E slides (Supplementary Fig. 1d), slides were tiled into 299-by-299-pixel pieces and were packaged into one TFrecords file for each set after color normalization. All models were trained from scratch. Later, to validate the generalizability of the prediction models, an independent test set consisted of samples only from the CPTAC cohort were used to evaluate models that were trained and validated solely on samples from TCGA cohort (Supplementary Fig. 2a).

We developed a multi-resolution InceptionResnet-based(Szegedy et al., 2017) convolutional neural network architecture, Panoptes, to capture features of various sizes on the H&E slides, which resembles the reviewing strategy of human pathologists. Unlike the conventional CNN architecture, the input of Panoptes is a set of 3 tiles of the same region on the H&E slide instead of a single tile. The resolution of tiles in a set is 2.5X, 5X, and 10X so that the higher resolution tile covers one fourth of the region in the next lower resolution tile (Fig. 1c). Hence, each grid region at 2.5X resolution in Fig. 1c can ideally generate 16 tile sets provided none of the tiles contained more than 40% of background pixels. Each set of tiles were

converted into a single matrix as one sample. Panoptes has three InceptionResnet-based branches, each of which processed the tiles with a specific resolution of the same region simultaneously (Fig. 1d). These branches worked separately until the third-to-last layer of the architecture, where inputs from the three branches were concatenated, followed by a global average pooling layer and the final fully connected layer. This design enabled the branches to learn features of different scales. More abstract information from each branch was integrated only at higher levels. We attempted to add an additional 1-by-1 feature pooling convolutional layer before the global average pooling and introduced a fourth branch processing clinical features. The effectiveness and comparisons of these modifications are discussed in later sections. Compared to conventional CNN, the multi-resolution design of Panoptes can at the same time considers both macro tissue-level features and minute cellular-level features of the same region, and can therefore capture more comprehensive characteristics of the slides. Moreover, taking the grouped multi-resolution tile sets as input while having a single output and loss function preserves the original positional information, which makes Panoptes distinct from the simply joining decisions from three separate models trained on tiles of three resolutions. To find the best performing model for different prediction tasks, we tried four different Panoptes architectures with and without the clinical feature branch, two types of InceptionResnet, and three types of Inception in this study. InceptionResnet and Inception models were trained on single resolution 10X tiles. Among all the statistical metrics calculated, we used area under the receiver operating characteristic curve (AUROC) of the test sets as the major metrics to evaluate the performance of the models, which is the typical standard in the machine learning field. Precision, recall, sensitivity, specificity, and F1 scores were also used to evaluate imbalanced prediction tasks. Per-patient level prediction was obtained by taking the mean of the predicted probability (prediction score) of all tiles belonging to the same patient. The AUROC was then calculated by taking each patient as one sample point. For Panoptes models, one set of grouped tiles was counted as a single tile for the metric calculation purpose since the output from the model was only one prediction score for each set.

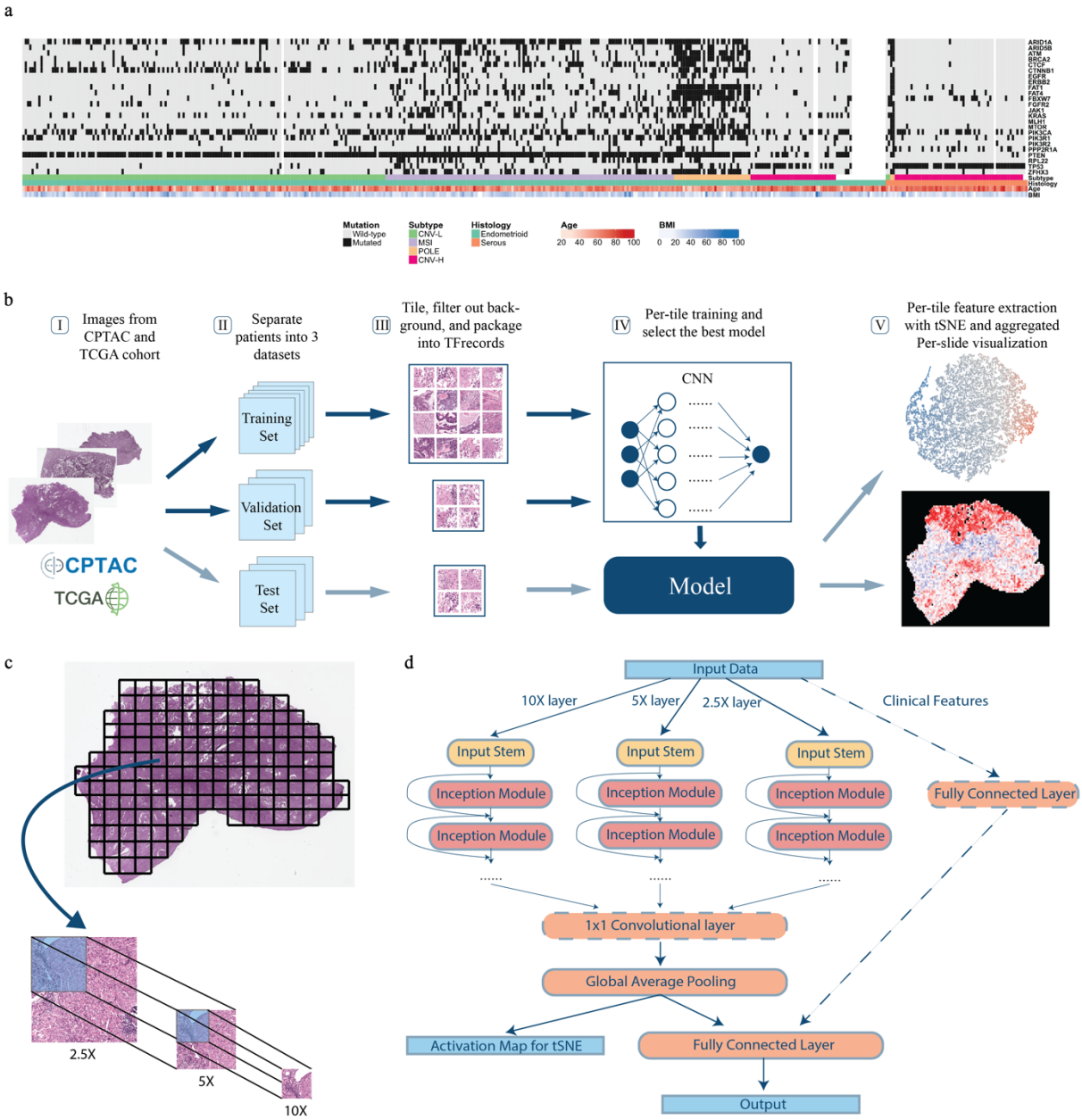


Fig.1 | Workflow and Panoptes architecture. a, Patients in the cohorts with feature annotations. **b**, Overall workflow. (**b, I**), H&E slide images of endometrial cancers from CPTAC and TCGA were downloaded; (**b, II**), slides were separated at per-patient level into a training (80%), a validation (10%), and a test set (10%); (**b,III**), slides were cut into 299x299-pixel tiles excluding background and contaminants and qualified tiles were packaged into TFrecord files for each set; (**b, IV**), training and validation sets were used to train the convolutional neural networks and the testing set was used to evaluate trained models; (**b, V**), activation maps of test set tiles were extracted and dimensionally reduced by tSNE to visualize features while the per-tile predictions were aggregated back into intact slides. **c**, Slides were cut into paired tile sets at 2.5X, 5X, and 10X equivalent resolution of the same region to prepare for Panoptes. **d**, Panoptes architecture with optional 1X1 convolutional layer and clinical features branch.

Multi-resolution deep-learning architectures achieved better predictive performance on histopathology images. We trained models to predict histological subtypes, CNV-H subtype from the entire cohort and the endometrioid patients, MSI-High subtype, and the mutation status of 18 endometrial-carcinoma-related genes. We applied five baseline models (InceptionV1, InceptionV2, InceptionV3, InceptionResnetV1, and InceptionResnetV2) and four versions of multi-resolution models (Panoptes1-4) on all of the tasks. The same data splits were used for all the models of the same predictive tasks in order to have fair comparisons among different architectured models. The best performing architectures for each of the prediction tasks and their corresponding AUROCs with 95% confidence intervals (CI) are shown in Table 1. Tasks with per-patient AUROC less than 0.6 were not listed. We performed 1-tail Wilcoxon tests on prediction scores between positively and negatively labeled tiles for the results in Table 1 and they all showed significant differences (Fig. 2a). Therefore, the prediction scores of true-label-positive tiles were significantly higher than those of true-label-negative tiles, demonstrating that these models were able to distinguish tiles in the test sets. The complete AUROC performance of all these trials are shown in Supplementary Fig. 2b-c. ROC curve examples of the top four prediction tasks are shown in Fig. 2b-c.

Based on the AUROC scores, we observed that Panoptes models were the best architectures in the top five prediction tasks (Table1). It is also clear that Panoptes performed better than Inception and InceptionResnet models for most of the tasks (Supplement Fig. 2). To validate that Panoptes performed better than InceptionResnet, we conducted 1-tail t-test on AUROC performance of the top eight prediction tasks between the Panoptes models and their corresponding InceptionResnet models. Panoptes2, which was the best Panoptes architecture in most of the tasks, showed a significantly higher AUROC than the corresponding InceptionResnet2 in six out of eight prediction tasks at per-patient level and seven out of eight at per-tile level (Fig. 2d-e). Similarly, Panoptes1 had a significantly higher AUROC than InceptionResnet1 in five out of eight prediction tasks at per-patient level and seven out of eight at per-tile level (Supplementary Fig. 3a-b).

To evaluate the effectiveness of adding an additional 1-by-1 convolutional layer between concatenation of branches and the global average pooling, we performed a 1-tail t-test between Panoptes1 and Panoptes3 as well as Panoptes2 and Panoptes4. However, only two tasks at per-patient level and five tasks at per-tile level showed significant p-values between Panoptes2 and Panoptes4 (Supplementary Fig. 3c-d). Similar results were observed between Panoptes1 and Panoptes3, where only one per-patient level task and four per-tile level tasks showed a significant difference. By applying the same test to Panoptes with and without clinical feature branch models, most of the tasks were not statistically significant with an example of Panoptes2 having two significant tasks at per-patient level and four at per-tile level (Supplementary Fig. 3e-f). In summary, our multi-resolution architectures Panoptes outperformed InceptionResnet in analyzing endometrial cancer H&E slides in various prediction tasks. The effectiveness of the additional convolutional layer and the integration of patients' age and body mass index (BMI) through a fourth branch was not found to be significant, however.

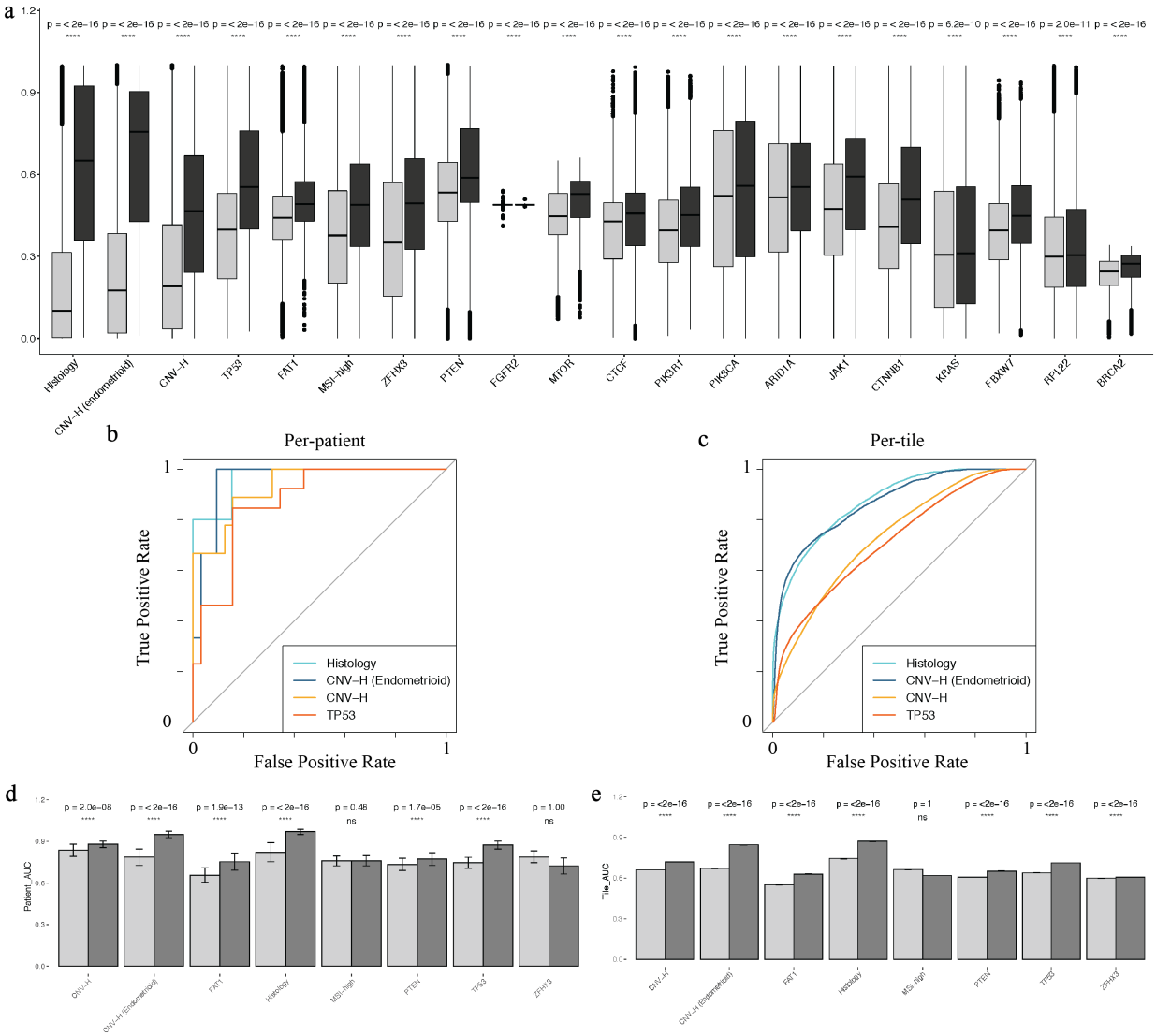


Fig.2 | Prediction tasks were statistically successful with promising results and Panoptes outcompeted baselines in most of top-performing prediction tasks. a, Predicted positive probability of tiles with 1-tail Wilcoxon test between true label positive and negative groups (black: true label positive tiles; grey: true label negative tiles) from models in **Table 1**. **b, c**, ROC curves at per-patient (**b**) and per-tile (**c**) level associated with the top four tasks in **a**. **d, e**, Bootstrapped per-patient (**d**) and per-tile (**e**) AUROC of InceptionResnetV2 (light) and Panoptes2 (dark) of top eight tasks in a with 1-tail t-test.

Accurate predictions of histological and molecular subtypes. Panoptes2 achieved a 0.969 (CI: 0.905-1) per-patient level AUROC in classifying samples into endometrioid or serous histological subtypes with an F1 score of 0.75. The precision was 1 and the recall was 0.6 respectively at per-patient level. Panoptes models were in the leading positions followed by InceptionV3 and InceptionV2, all of which had per-patient AUROC above 0.9. For molecular subtyping tasks, we applied all architectures on four binary tasks, each aimed at predicting one molecular subtype versus all others. Panoptes1 achieved a per-patient AUROC of 0.934 (CI: 0.851-1) in predicting CNV-H while all other Panoptes models achieved an AUROC above 0.88, outcompeting the baseline models by 5.8% to 23.3%. The best F1 score was 0.8 with precision of 0.727 and recall of 0.889 respectively. This model also achieved sensitivity of 0.889 and specificity of 0.906 when using 0.5 as the cutoff point of prediction scores. Apart from CNV-H, we also trained models classifying another molecular subtype, MSI-High, with a best per-patient AUROC of 0.827 (CI: 0.705-0.948) and F1 score of 0.615.

Although most CNV-H cases are of serous subtype, a portion of high-grade endometrioid cancers are also classified as CNV-H. To further assess whether machine learning models could capture the heterogeneity within this histological subtype, we trained models to predict CNV-H status in endometrioid samples. The Panoptes1 architecture was able to achieve a per-patient AUROC of 0.958 (CI: 0.886-1) and F1 score of 0.667 on this task, suggesting that the model utilized features that were not strongly associated with histological subtype to predict molecular subtype. All Panoptes models also outcompeted baseline models in this task. In addition, we trained models to predict mutation status of 18 driver genes. Panoptes2 was able to predict a *TP53* mutation with a per-patient AUROC of 0.873 (CI: 0.768-0.977) and F1 score of 0.56. *FAT1* mutation was predicted using Panoptes2 (with clinical feature branch) with a per-patient AUROC of 0.835 (CI: 0.666-1) and F1 score of 0.545. Other gene mutations, including *ZFHX3*, *PTEN*, *FGFR2*, *MTOR*, *CTCF*, and *PIK3R1*, were also predicted with a per-patient AUROC above 0.7. A table showing the full statistical metrics for all the prediction models can be found in the supplementary files.

Table 1 | AUROCs of the best models for each task with 95% confidence intervals (CIs).

	Best Architecture	Per-patient AUROC	Per-tile AUROC
Histology	Panoptes2	**0.969 (0.905-1)**	**0.870 (0.866-0.874)**
CNV-H from endometrioid	Panoptes1	**0.958 (0.886-1)**	**0.864 (0.859-0.870)**
CNV-H	Panoptes4	**0.934 (0.851-1)**	0.731 (0.728-0.734)
TP53	Panoptes2	**0.873 (0.768-0.977)**	0.713 (0.709-0.717)
FAT1	Panoptes2 with clinical	**0.835 (0.666-1)**	0.639 (0.635-0.642)
MSI-High	InceptionResnetV1	**0.827 (0.705-0.948)**	0.638 (0.635-0.641)
ZFHX3	InceptionResnetV1	**0.824 (0.689-0.959)**	0.637 (0.634-0.640)
PTEN	InceptionV2	**0.781 (0.579-0.984)**	0.623 (0.620-0.627)
FGFR2	Panoptes4 with clinical	**0.755 (0.540-0.970)**	0.550 (0.545-0.554)
MTOR	Panoptes1	0.724 (0.496-0.951)	0.674 (0.670-0.678)
CTCF	Panoptes4	0.724 (0.518-0.931)	0.571 (0.568-0.575)
PIK3R1	InceptionResnetV1	0.702 (0.524-0.880)	0.596 (0.593-0.599)
PIK3CA	Panoptes4	0.689 (0.532-0.847)	0.526 (0.523-0.530)
ARID1A	InceptionResnetV2	0.683 (0.513-0.853)	0.542 (0.538-0.545)
JAK1	Panoptes2 with clinical	0.662 (0.410-0.940)	0.612 (0.605-0.618)
CTNNB1	InceptionResnetV2	0.648 (0.439-0.858)	0.619 (0.616-0.622)
KRAS	Panoptes2 with clinical	0.638 (0.404-0.871)	0.515 (0.510-0.519)
FBXW7	InceptionV3	0.629 (0.366-0.892)	0.606 (0.602-0.609)
RPL22	InceptionV3	0.632 (0.395-0.868)	0.517 (0.512-0.522)
BRCA2	InceptionResnetV1	0.613 (0.318-0.908)	0.624 (0.620-0.629)

Feature extraction and whole-slide visualization revealed correlations and differences between histological and molecular features. To visualize and evaluate features learned by the models for each task, we extracted the activation maps before the final fully connected layer of the test set tiles. 20000 tiles' activation maps were then randomly sampled for each task. These activation maps were dimensionally reduced and displayed on 2D tSNE plots, where each dot represents a sampled tile and was colored according to the positive prediction scores. As we expected, tiles were generally clustered by their predicted groups. By replacing dots with the original input tiles of different resolutions, we were able to discover features that correlated with the predictions corresponding to the specific histological or molecular classification task. For example, features of predicted histologically serous and endometrioid were drastically different (Fig. 3a). In the cluster with high prediction scores of serous subtype, we observed typical serous carcinoma features, such as high nuclear grade, papillary growth pattern, elevated mitotic activity, and slit-like spaces. Tiles in the cluster of predicted endometrioid cases showed low nuclear grade, glandular growth pattern, cribriform architecture, and squamous differentiation. Myometrium and other non-tumor tissue tiles were located in the middle of the tSNE plot with prediction scores between 0.4 and 0.6 (Fig. 3a). These observations suggested that our models were able to focus on the tumor regions of H&E slides and make histological subtype predictions based on features that were also recognized by human experts in pathology.

The features learned by molecular subtype prediction models were also revealed with the same feature extraction method. We noticed that in the CNV-H prediction model, two distinct subgroups were recognized in the predicted CNV-H cluster, associated with histological serous and high grade endometrioid subtypes, respectively (Fig. 3b). The predicted-CNV-H serous tiles mostly showed high nuclear grade, gland formation, and elevated mitotic activity, while the predicted-CNV-H high grade endometrioid tiles exhibited solid growth pattern and focal glandular differentiation. On the other hand, in the non-CNV-H cluster, tiles were mostly low-grade endometrioid carcinoma with low nuclear grade, gland formation, and squamous differentiation (Fig. 3b). To confirm that the tumor grade was the major factor to distinguish CNV-H molecular subtype in endometrioid samples, we unveiled the features learned by the CNV-H prediction model trained only on endometrioid images (Fig. 3c). As we expected, high-grade endometrioid carcinoma tiles were observed mostly in the CNV-H cluster, leaving the low-grade tiles in the non-CNV-H cluster. In both of these CNV-H models, the ambiguous regions were mostly occupied by non-tumor tissue. We also visualized the major pattern learned by the model to distinguish MSI-H subtype images from others (Fig. 3d). Tiles in the MSI-H cluster were mostly low grade endometrioid carcinomas with gland formation, tumor infiltrating lymphocytes, and peritumoral lymphocytes, consistent with the observation that heavy mutation load of MSI-H tumors lead to high immunogenicity and a host immune response[27,28].

In addition to the subtypes, patterns related to some mutations were also revealed. A *PTEN*-mutated cluster mostly contained tiles of low grade endometrioid carcinomas with gland formation and low nuclear grade (Supplementary Fig. 4a) while *TP53*-mutated tiles were generally serous carcinomas with high nuclear grade and abundant tufting and budding (Supplementary Fig. 4b). Furthermore, low grade endometrioid carcinoma tiles with gland formation, low nuclear grade, and abundant tumor infiltrating lymphocytes were present in the *ZFHX3*-mutated cluster while those with much less lymphocytes were in the wild type cluster (Supplementary Fig. 4c). High grade endometrioid carcinoma tiles with diffuse solid growth and low nuclear grade were depicted in *FAT1*-mutated cluster while low grade endometrioid carcinomas with gland formation, low nuclear grade, and cribriform architecture were in the wild type cluster (Supplementary Fig. 4d). These findings may result from the correlation between mutation status and histological or molecular subtypes described above, as *PTEN* and *TP53* mutations were mainly found in endometrioid and serous subtypes, respectively, while *ZFHX3*

and *FAT1* mutation status showed correlation with the heavily mutated MSI-H and *POLE* molecular subtypes(Supplementary Fig. 1c).

Additionally, we were also interested in visualizing distribution of features on the whole slide level. Prediction of tiles from the test sets were aggregated back to the size of original slides in the form of heatmaps, where hotter tiles corresponded to higher positive prediction scores. Whole slide visualization revealed that our models tended to have extreme prediction scores on tumor regions instead of non-tumor tissues such as myometrium (Supplementary Fig. 5).The first slide in Fig. 4 was from an endometrioid and CNV-H case while the second slide was from a serous and CNV-H case. Models correctly predicted both tasks for the two slides. By comparing the prediction of histological subtypes and CNV-H, we found that the models were focusing on different yet related features in these 2 prediction tasks. In the first slide, the areas predicted to be endometrioid were largely classified as CNV-H while in the second slide, most areas predicted as serous were also classified as CNV-H. Although most CNV-H samples were histologically serous, our models were capable of learning an additional set of features. In other words, the prediction of CNV-H in our models were not necessarily based on features of histological subtypes.

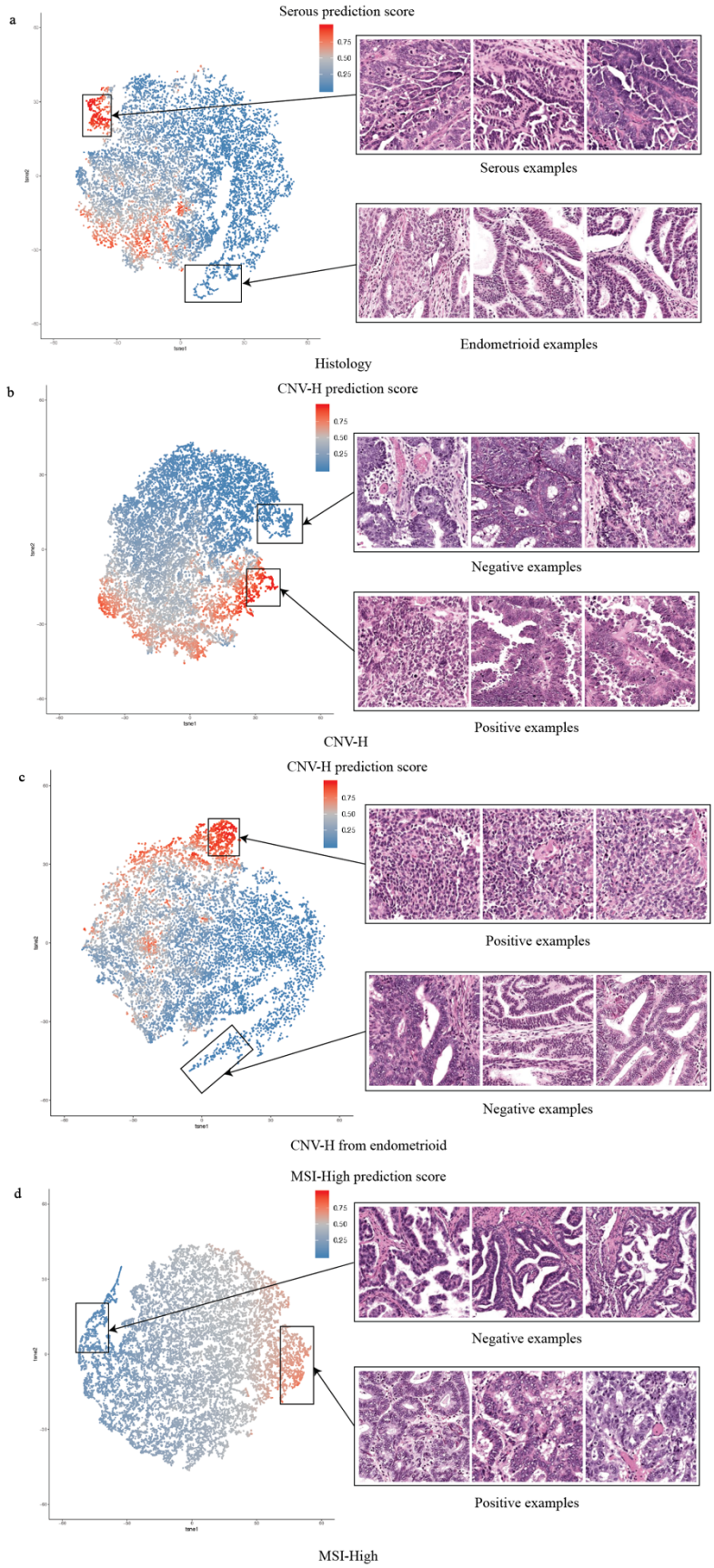


Fig.3 | Extraction and visualization of features learned by the models with tSNE. Each point represents a tile and is colored according to its corresponding positive prediction score. **a**, Histologically serous and endometrioid features from a Panoptes1 model. **b**, CNV-H positive and negative features from a Panoptes4 model. **c**, CNV-H positive and negative features in the histologically endometrioid samples from a Panoptes1 model. **d**, MSI-High positive and negative features in the histologically endometrioid samples from a Panoptes3 with clinical features model.

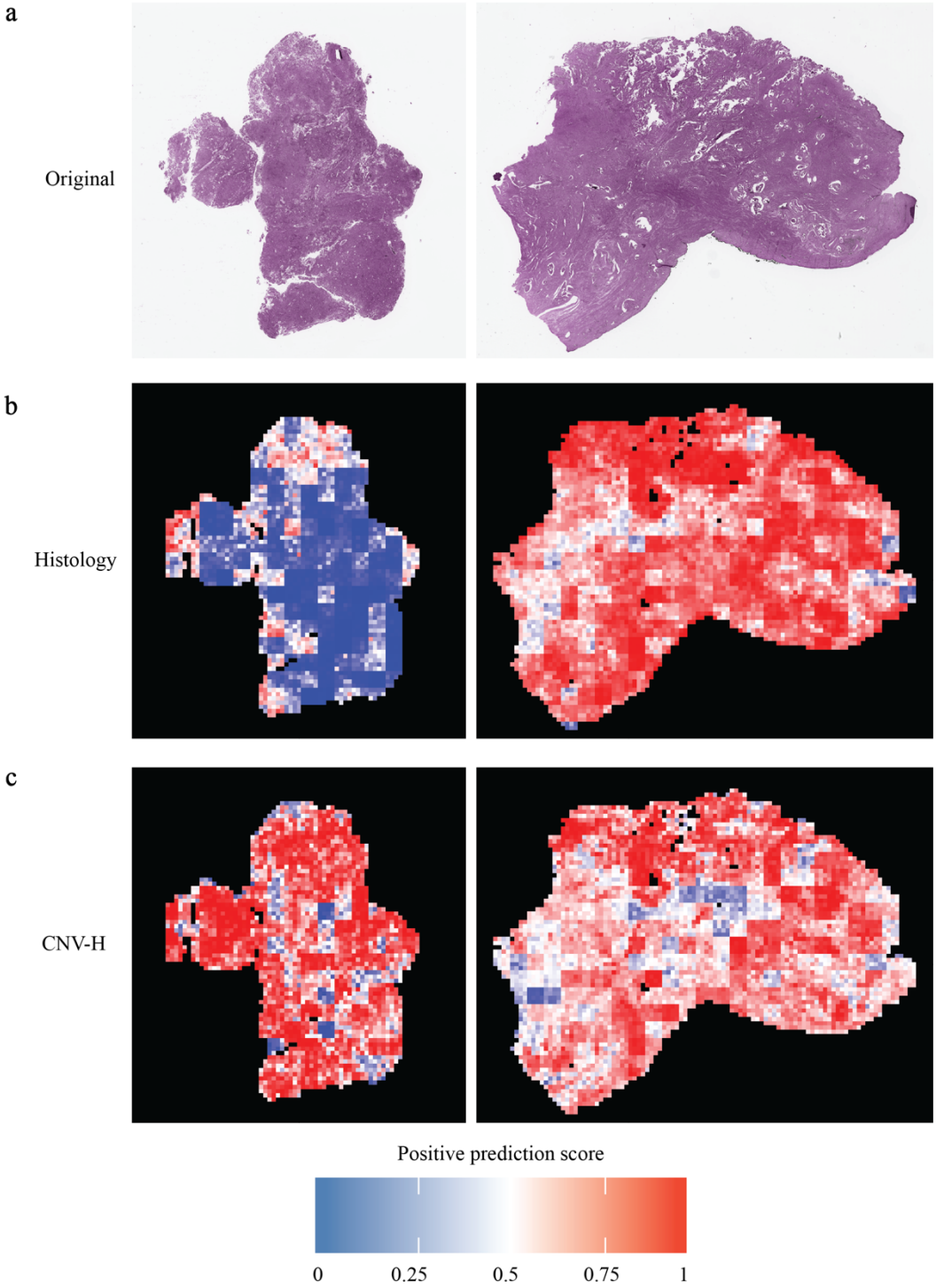


Fig. 4 | Whole slide predictions showing features of histological subtype and CNV-H are distinct but correlated. a, The first example slide is from a CNV-H but histologically endometrioid case while the second example slide is from a CNV-H and serous tumor. **b**, Whole slide histology prediction of examples in **a** Panoptes2 model with hotter regions being predictive more of serous while cooler regions were more endometrioid. **c**, Whole slide CNV-H prediction of examples in **a** Panoptes1 (first example) and Panoptes4 (second example) models with hotter regions being more predictive of CNV-H.

Generalizability and potential clinical capability of the models was illustrated by testing with samples from an independent cohort. To prove that our models were generalizable and the predictions results were consistent and not overfitted, we repeated the training and testing of models for all the tasks using independent cohort data split and compared statistical metrics with the previous mixed data split results. In these independent data split trials, models were trained and validated only on the data from TCGA at a 9:1 split ratio. Slides from CPTAC served as an independent test set for all the prediction tasks. Therefore, the size of training and validation sets of these trials were smaller and less diversified than the mixed data slit trials. All other hyperparameters remained the same for the training, validation, and testing workflow. The AUROC of CPTAC independent test set indicated that Panoptes-based models still had better performance than the baseline models in general (Supplementary Fig. 2d-e). The best performing models based on independent test set were compared side-by-side with the best models in mixed random split trials (Fig. 5). A Panoptes4 model achieved an AUROC of 0.962 (CI: 0.926-0.999) with F1 score of 0.696 at per-patient level, which were similar to the best model trained and tested on the mixed data split. In the CNV-H molecular subtype prediction task, a Panoptes3 model showed an AUROC of 0.87 (CI: 0.753-0.987) with F1 score of 0.667 at per-patient level. Slightly lower performances were also observed in prediction tasks using independent data split at per-patient level, including CNVH in endometrioid, MSI-high, *TP53*, and *FAT1*. However, higher statistical metrics were observed in some prediction tasks, such as *PTEN*, *KRAS*, *BRCA2*, and *CTNNB1*. Interestingly, even though the per-patient level metrics were lower in the independent data split trials than in the mixed data split trials for some prediction tasks (CNV-H, *TP53*, *CTCF*), their per-tile level metrics were higher. In addition, we compared Panoptes-based models' performance side-by-side in mixed random split trials and CPTAC independent test set trials and the results were similar to the best performing models comparison (Supplementary Fig. 6). The full table of statistical metrics of the test set in the independent cohort split trials can be found in the supplementary files.

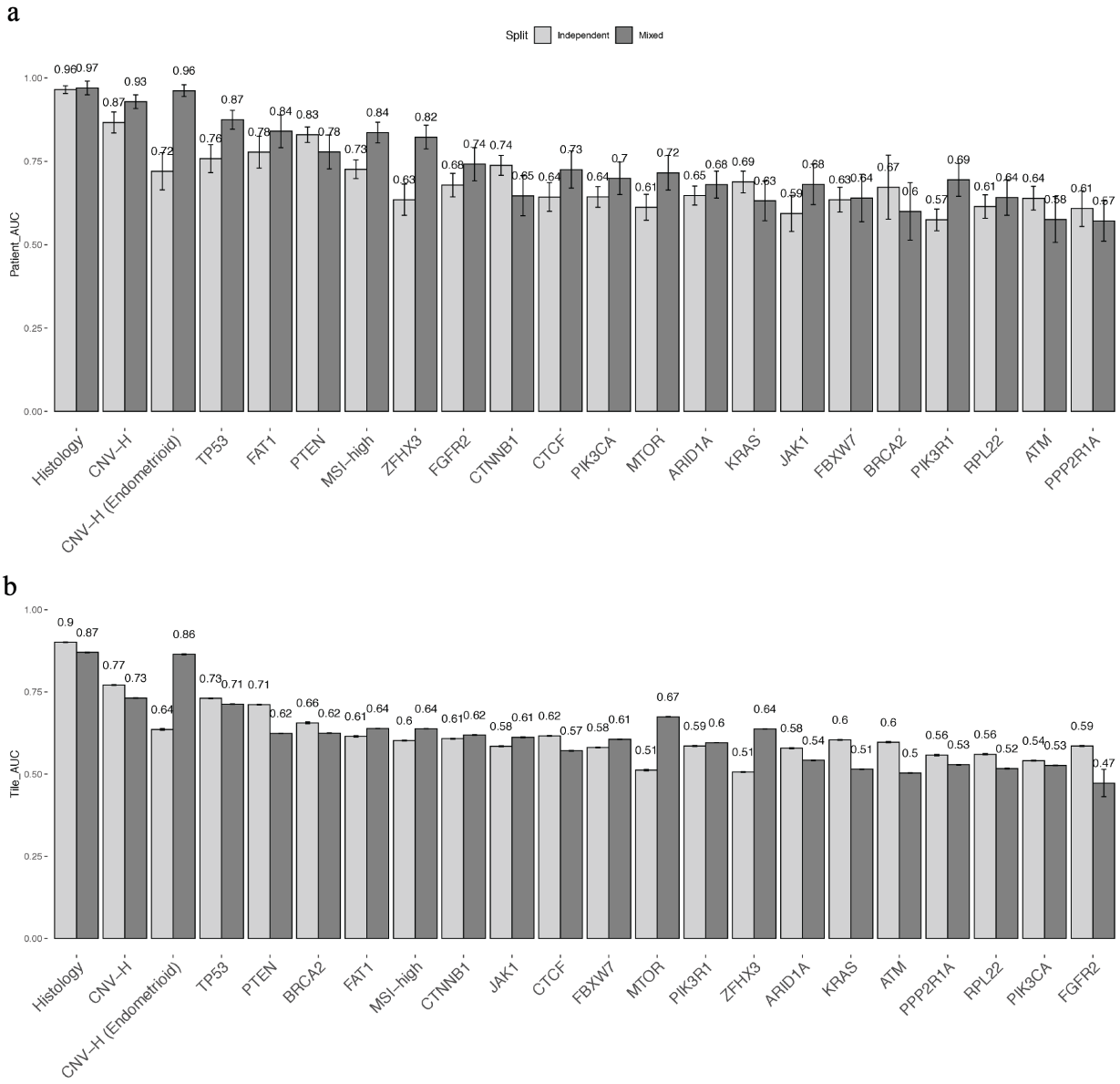


Fig. 5 | Side-by-side comparisons of AUROC between the best models in mixed random split trials and independent cohort split trials revealed the generalizability of the models. Per-patient (**a**) and per-tile (**b**) level AUROC of the best performing models in each task with mixed random data split (dark grey) and the cohort independent data split (light grey).

Discussion

Our study introduced a novel multi-resolution InceptionResnet-based convolutional neural network architecture, Panoptes, which was able to accurately predict endometrial cancer histological and molecular subtypes as well as mutation status of critical genes based on H&E slides. The AUROC of classifying endometrioid and serous histological subtypes by our best architecture model was 0.969 (CI: 0.905-1). Moreover, the models can distinguish the most lethal molecular subtype, CNV-H, with exceptionally high accuracy (AUROC 0.934). It is worth noting that our models can also precisely identify the CNV-H samples from a histologically endometrioid carcinoma (AUROC 0.958), which is one of the more controversial and complex patient subgroups in endometrial cancer subtyping. In addition to the CNV-H, we were also able to predict other molecular features with acceptable performance, which are currently not possible for pathologists to determine without ancillary studies such as sequencing or immunohistochemistry. These include the DNA-mismatch repair deficiency-related MSI-high molecular subtype (AUROC 0.827), the mutation the CNV-H signature gene *TP53* (AUROC 0.873) as well as *PTEN* (AUROC 0.781), *FAT1* (AUROC 0.835) and *ZFHX3* (AUROC 0.824). Statistical analyses proved the success of our prediction tasks. In addition, we tested and showed that our multi-resolution Panoptes-based models performed significantly better than InceptionResnet-based models in most of our prediction tasks. We attempted two modifications to Panoptes, including an additional convolutional layer and integration of clinical features, but failed to observe significant improvement in performance. By extracting and clustering abstract representation of tiles constructed by the model, we discovered critical features to distinguish subtypes and mutations, particularly the tumor grade in determining CNV-H cases from non-CNV-H samples. We justified the generalizability and potential clinical applicability of our models by training and validating on samples from TCGA while testing solely on an independent cohort with samples from CPTAC. Although slightly lower performances were observed in the CPTAC independent testing trials for some prediction tasks, we believe that it was mostly likely caused by smaller and less diversified TCGA-only training set.

Examining H&E slides is still currently the most widely used techniques for pathologists to confirm endometrial cancer histological subtypes in the clinical setting. Our models showed great potential in assisting pathologists making decisions and improving diagnostic accuracy. Given most H&E slides can be tiled into less than 5000 tile-sets (Supplementary Fig. 1e-g), with a processing speed of 22 tile-sets per second (1310 tile-sets per minute) on a Quadro P6000 GPU, our models can analyze a slide within 4 minutes. This means that these models can work simultaneously with pathologists to serve as references. We have shown that the model utilized human interpretable features to perform histological and molecular classification tasks. With whole slide visualization, the reassembled per tile predictions can provide a thorough examination of the H&E slide and a detailed layer containing potential hotspot feature, which may also include regions that could possibly be neglected by pathologists. However, due to the time-consuming H&E slide scanning and tiling processes, multiple optimizations need to be implemented before the system could be deployed in practice.

Both histological and molecular features' labels of TCGA and CPTAC samples have been validated by many scientists and clinicians before and after the publication of their studies However, as tile labels were assigned at per-patient level, within-slide heterogeneity would still lead to noise in the true labels, such that features in a local region may not match the characteristics of the assigned classification. Therefore, we believe that the per-patient metrics are more accurate than per-tile metrics in terms of accessing a models performance. The performance can be further improved if more detailed annotations existed on the slides. From the visualization results, we noticed that our models were more likely to give non-tumor tissue tiles ambiguous prediction scores (0.4-0.6). Therefore, building a segmentation model to exclude these irrelevant non-tumor tissue, such as myometrium, may also significantly enhance the overall performance of our models.

Although the TCGA and CPTAC datasets cover a variety of endometrial carcinoma samples, it may not reflect the full pathological diversity of endometrial cancer. In this study, we tried two data split criteria, mixed random split and cohort independent split, to create training, validation, and test sets. We adopted the random split of the mixed TCGA and CPTAC dataset to train models that could achieve the best performance as the models can learn from both datasets instead of TCGA or CPTAC specific features. From the other perspective, we used cohort independent data split, namely training and validating on TCGA samples while testing on CPTAC samples to justify the generalizability and indicate potential clinical capability of our models. Additionally, multiple trials took place to find the best normalization method, combination of hyperparameters, and architecture designs.

Overall, we demonstrated that our multi-resolution convolutional neural network architecture, Panoptes, can be a practical tool to assist pathologists classifying endometrial cancer histological subtypes and, more importantly, to provide additional information about patients' molecular subtypes and mutation status in a much more rapid fashion and without the need for sequencing. In addition to per-patient level prediction, the model would also be able to highlight regions with human interpretable features on the slide. Moreover, it remains possible that our models have learned visual patterns which correlated with molecular features that were not previously annotated by human experts and requires further investigation. From another perspective, these novel patterns from the H&E slides may be incorporated into the current standards of histological pathology and contribute to improved prognosis and treatment of endometrial carcinoma in the future.

Our future plan includes refining the Panoptes architecture, particularly to determine an effective way to integrate clinical features into the imaging prediction branches to improve the overall performance. Quantification of features could also be added to the Panoptes. We would also work on training our existing models with slides labeled with more detail and new datasets that cover more heterogeneity and diversity of endometrial cancer in order to make the models more robust and generalizable. Predicting other molecular subtypes and mutations, such as *POLE*, *CTNNB1*, and *JAK1*, which did not have a well-performing model in this study, will be possible once more data are available. In addition to the currently available user interface of Panoptes, we plan to develop a more advanced Graphical User Interface (GUI) that includes all the trained models and outputs visualization and prediction in a fast and user-friendly way, which we are hoping to be deployed and tested in a pathologist's clinical practice. We would try to train Panoptes-based models to predict features in other types of cancers, such as glioblastoma, melanoma, and lung carcinoma, and it would be very interesting to see how Panoptes performs and what features it captures in these new tasks.

Methods

Data Acquisition and Summary

We used samples from 2 datasets, The Cancer Genome Atlas (TCGA) and CPTAC (Clinical Proteomic Tumor Analysis Consortium). 392 diagnostic slides from 361 Uterine Corpus Endometrial Carcinoma (UCEC) patients in TCGA cohort were downloaded from the NCI-GDC Data Portal. These samples were published in the TCGA pan-cancer atlas. Demographic, genomic, and other clinical features associated with these samples were downloaded from the cBioPortal and the original TCGA UCEC paper supplements[15]. 107 diagnostic slides from 98 Uterine Corpus Endometrial Carcinoma (UCEC) patients in CPTAC cohort were downloaded from The Cancer Imaging Archive (TCIA). Demographic, genomic, and other clinical features of these patients were published in the CPTAC UCEC paper[25]. The composition of patients with different features of interests are shown in Fig. 1a. Most of the patients in our cohort have only 1 diagnostic slide (Supplementary Fig. 1b).

H&E Images Preparation

Digital histopathologic images were in SVS format, which were tuples of the same images with 3 or 4 different resolution levels. Slides from the TCGA cohort were scanned with a maximum resolution of 40x while those from the CPTAC cohort were at 20x maximum resolution. A Python package, Openslide, was used to maneuver the SVS files. Due to the extremely large size of these images (Supplementary Fig. 1d), they were cut into small tiles in order to be fed into the training pipeline. Multi-threading was used to accelerate this process. Tiles were cut at 10x, 5x, and 2.5x equivalent resolutions for both cohort and algorithm was used to exclude tiles with more than 40% pixels of white background and irrelevant contaminants (Supplementary Fig. 1e-g). Stain colors of the useful tiles were normalized using the Vahadane's method during this process[29]. For each of the tasks, the labels were one-hot encoded at per-tile level. The datasets were separated into training, validation, and testing sets at per-patient level with a ratio of 8:1:1. To take advantage of the Tensorflow API and accelerate the training and testing process, tiles were loaded and saved into a single TFrecords file for each set.

Baseline Models

InceptionV1, InceptionV2, InceptionV3, InceptionResnetV1, and InceptionResnetV2 architecture were trained from scratch and used as the baseline models. InceptionResnets are enhanced architectures of Inceptions with residual connections and a previous study has shown that they are performed generally better than Inceptions in imaging prediction tasks[26] . The auxiliary classifiers of these architectures were opened. We did not modify any part of the backbone of these architectures. Tiles with 10x resolution were input and we used back-propagation, softmax cross entropy loss weighed by training data composition, and Adam optimization algorithm in the training workflow. Here, each single tile image with a label is considered 1 sample. Batch sizes were set to 64 with an initial learning rate of 0.0001 and a drop-out keep rate of 0.3. We tested multiple combinations of hyperparameters and found that this one achieved optimal results for most tasks. The training jobs were run with no fixed epoch number. 100 batches of validation were carried out every 1000 iterations of training and when the training loss achieved a new minimum value after 30000 iterations of training. If the mean of these 100-batch validation loss achieved minimum, the model was saved as the temporary best performing model. The training process stopped when the validation loss did not decrease for at least 10000 iterations. This stopping criterion was only initiated after 100000 iterations of training.

Panoptes Models

We used 4 different Panoptes architectures with and without the integration of patients' BMI and age in a fourth branch. Panoptes1 has 3 branches based on InceptionResnet1 and Panoptes2

has 3 branches based on InceptionResnet2. The major difference of Panoptes3 to Panoptes1 and of Panoptes4 to Panoptes2 is the additional 1-by-1 convolutional layer between the concatenation of branches and the global average pooling. All of our Panoptes architectures were trained with randomly initialized network parameters with auxiliary classifiers opened on each branch. Unlike the baseline models, tiles of 10x, 5x, and 2.5x resolutions of the same region on the H&E slide with label were paired and considered as 1 sample as only 1 prediction score is associated with a multi-resolution matrix. Batch size was set to 24, which was the largest number that could fit in the memory of our GPUs. Optimization algorithm, weighted loss function, and other hyperparameters were the same as the baselines. In addition, we applied the same validation method to pick the best performing models and kept the same stopping criterion as the baselines.

Statistical Analyses

The performance was evaluated by applying the trained models to the test set. Each of the classification tasks has its own test set, which consists of slides from patients that had not been in the training or validation sets. Evaluation was performed at both per-patient level and per-tile level. Per-patient level metrics were obtained by taking the mean of all tiles' metrics that belonged to the same patient. For Panoptes model, a 3-multi-resolution-tile matrix is considered as 1 tile for statistical analyses. Receiver Operating Characteristic (ROC) curve, plotting true positive rate against false positive rate, and the area under the ROC curve (AUROC) were the major factors in evaluation. In addition, Precision Recall Curve (PRC), as well as average precision score (AUPR score), were used to determine the trade-off between false negative rate and false positive rate. We also used accuracy with softmax prediction score directly from the models. If the prediction score was greater than 0.5, it was counted as a positively predicted case. 95% Confidence intervals (CI) of AUROC, AUPR, and accuracy were estimated by the bootstrap method. Other statistical metrics, including sensitivity, specificity, precision, recall, F1 score, etc., were also generated and referred to evaluate the predictive models' performance. To further validate the effectiveness of the classification models, we did 1-tail Wilcoxon tests between positive and negative tiles in the test sets for each of the tasks. In order to compare performance between Panoptes models and the baselines, for each of the tasks with a patient level AUROC score greater than 0.75, we bootstrapped 50 times at an 80% sampling rate at both patient and tile level and calculated the AUROC for each of these sampled sets. Then, an unpaired 1-tail t-test between the AUROCs of Panoptes and its corresponding baseline model was performed. We performed a similar t-test between Panoptes with and without the additional convolutional layer as well as between Panoptes with and without the fourth branch of patients' BMI and age. Statistical analyses and plotting codes were written in R3.6 and Python3.

Feature Visualization Based on Tiles

For models with per-patient level AUROC above 0.75 of the test set, we randomly sampled 20000 tiles (tile sets for Panoptes) together with their feature maps before the last fully connected layer in the model, in which each tile or tile set is represented as a 1-dimensional vector. We then used tSNE with initial dimensions of 100 to reduce these 20000 vectors into 2-dimensional space where each point represents a tile or tile set. Generally, points clustered according to their predicted class. By replacing the points on tSNE plots with the original tiles, the features learned by the model for each of the specific class can be observed. We asked experienced pathologists to summarize the typical histological features in each of these clusters.

Whole Slide Prediction

We built an implementation pipeline that could apply trained models to whole H&E slides and output predictions as heatmaps. The heatmaps could be overlaid on the original slides, which

showed the prediction results of different areas. The maximum prediction resolution (each cell of the heatmap) is 299 by 299 pixel at 10x resolution level. Depending on the size of the H&E slides, the time of predicting an intact H&E slides can range from 2 to 40 minutes. The average speed of prediction with Panoptes models is 22 tile-sets per second, or 1310 tile-sets per minute.

Computational Resources and Code Availability

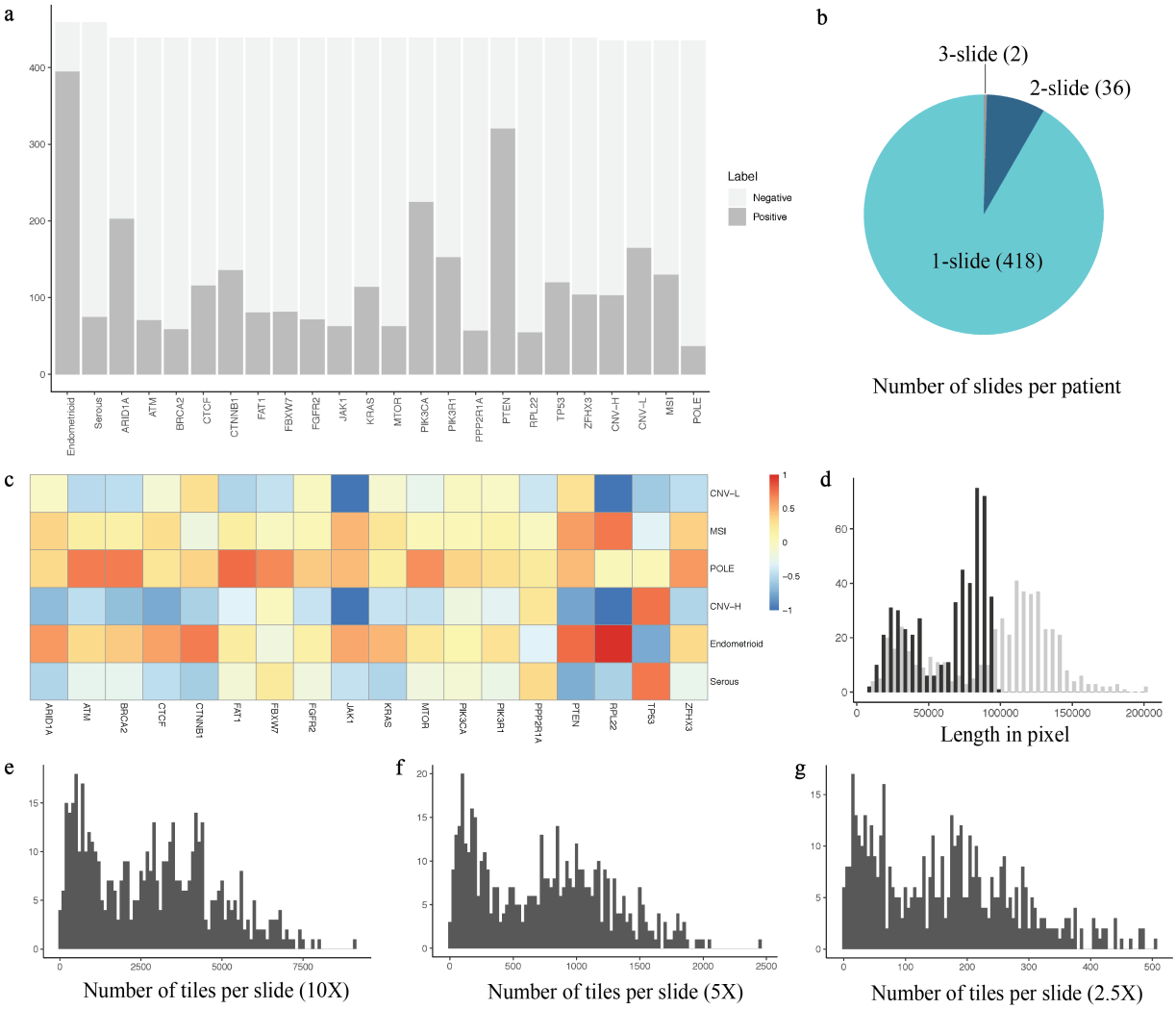
3 types of GPUs, NVIDIA Tesla P40, V100, and Quadro P6000 were used to train the models. We also used the NYU Prince HPC and NYU BigPurple HPC facilities for this project. The model training and testing codes were solely written in Python3 with Tensorflow 1.13 and they are compatible with Tensorflow 2. Statistical analyses and plotting codes were written in R3.6 and Python3. The analytic codes are available on GitHub in the following link: https://github.com/rhong3/CPTAC-UCEC. The Panoptes codes with user interface are available on Github in the following link: https://github.com/rhong3/Panoptes. The Panoptes Python3 package version is on PyPI with the following link: https://pypi.org/project/panoptes-he/

References

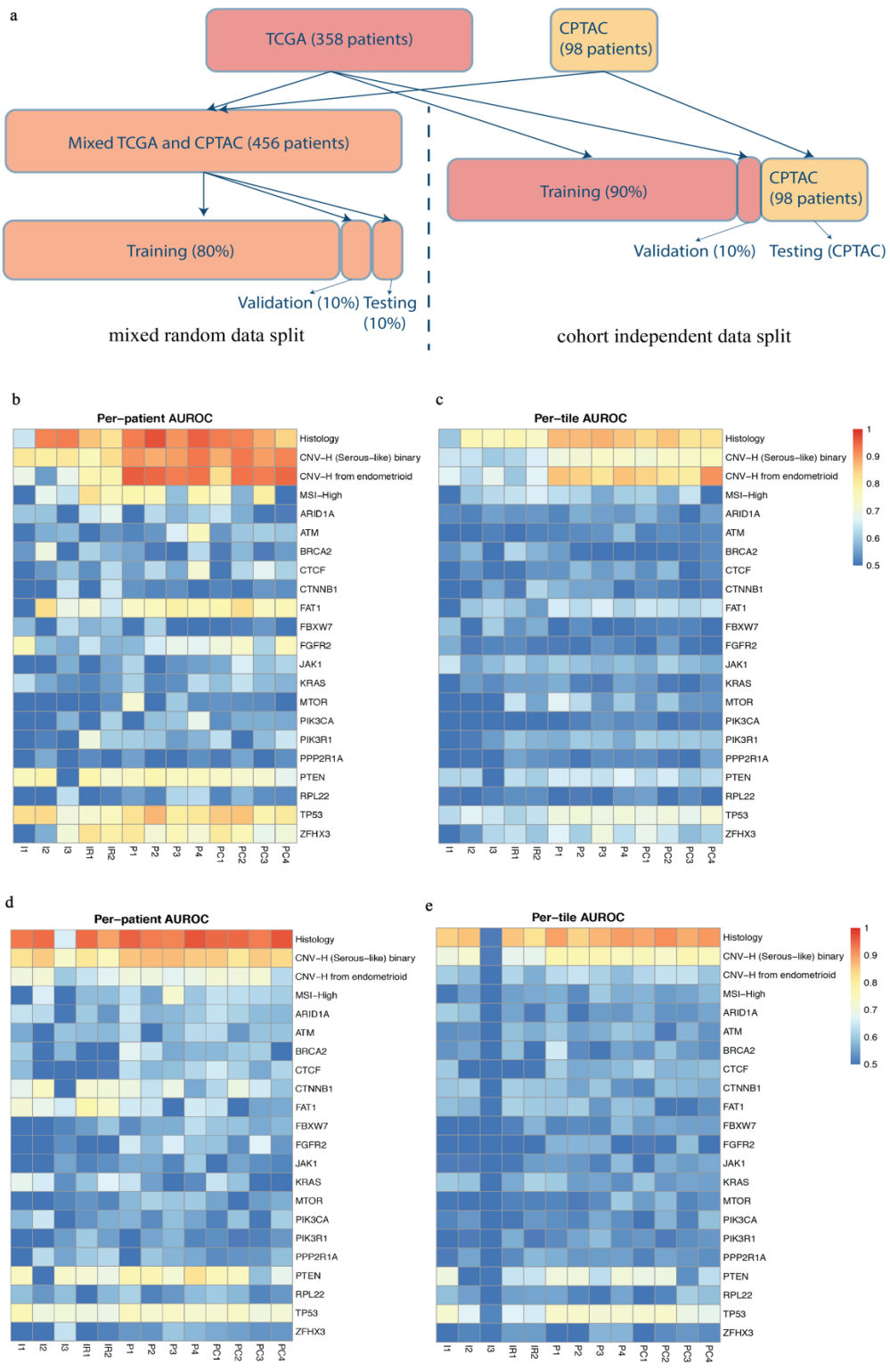
1. Amant, F. *et al.* Endometrial cancer. *Lancet* **366**, 491–505 (2005).
2. Morice, P., Leary, A., Creutzberg, C., Abu-Rustum, N. & Darai, E. Endometrial cancer. *Lancet* **387**, 1094–1108 (2016).
3. Burke, W. M. *et al.* Endometrial cancer: A review and current management strategies: Part I. *Gynecol. Oncol.* **134**, 385–392 (2014).
4. Burke, W. M. *et al.* Endometrial cancer: A review and current management strategies: Part II. *Gynecol. Oncol.* **134**, 393–402 (2014).
5. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA. Cancer J. Clin.* **66**, 7–30 (2016).
6. Home | American Cancer Society - Cancer Facts & Statistics. Available at: https://cancerstatisticscenter.cancer.org/#!/. (Accessed: 6th January 2020)
7. Bokhman, J. V. Two pathogenetic types of endometrial carcinoma. *Gynecol. Oncol.* **15**, 10–17 (1983).
8. Murali, R., Soslow, R. A. & Weigelt, B. Classification of endometrial carcinoma: more than two types. *Lancet Oncol.* **15**, e268–e278 (2014).
9. Frumovitz, M. *et al.* Predictors of final histology in patients with endometrial cancer. *Gynecol. Oncol.* **95**, 463–468 (2004).
10. Darvishian, F. *et al.* Serous Endometrial Cancers That Mimic Endometrioid Adenocarcinomas. *Am. J. Surg. Pathol.* **28**, 1568–1578 (2004).
11. Murray, S. K., Young, R. H. & Scully, R. E. Uterine Endometrioid Carcinoma with Small Nonvillous Papillae: An Analysis of 26 Cases of a Favorable-Prognosis Tumor To Be Distinguished from Serous Carcinoma. *Int. J. Surg. Pathol.* **8**, 279–289 (2000).
12. Murali, R. *et al.* High-grade Endometrial Carcinomas: Morphologic and Immunohistochemical Features, Diagnostic Challenges and Recommendations. *Int. J. Gynecol. Pathol.* **38**, S40–S63 (2019).
13. Brinton, L. A. *et al.* Etiologic heterogeneity in endometrial cancer: Evidence from a Gynecologic Oncology Group trial. *Gynecol. Oncol.* **129**, 277–284 (2013).
14. Zannoni, G. F. *et al.* Does high-grade endometrioid carcinoma (grade 3 FIGO) belong to type I or type II endometrial cancer? A clinical–pathological and immunohistochemical study. *Virchows Arch.* **457**, 27–34 (2010).
15. Getz, G. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
16. Bell, D. W. Novel genetic targets in endometrial cancer. *Expert Opin. Ther. Targets* **18**, 725–730 (2014).
17. Liang, S. & Lu, X. Research on the Inhibitory Effect of FAT-1 on Endometrial Cancer Cell Proliferation. *Am. J. Pharm* **37**, 903–910 (2018).
18. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* **33**, 170–175 (2016).
19. Louis, D. N. *et al.* Computational Pathology: A Path Ahead. *Arch. Pathol. Lab. Med.* **140**, 41–50 (2016).
20. Komura, D. & Ishikawa, S. Machine Learning Methods for Histopathological Image Analysis. *Comput. Struct. Biotechnol. J.* **16**, 34–42 (2018).
21. Nawaz, S. & Yuan, Y. Computational pathology: Exploring the spatial dimension of tumor ecology. *Cancer Lett.* **380**, 296–303 (2016).
22. Cooper, L. A. *et al.* PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective. *J. Pathol.* **244**, 512–524 (2018).
23. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
24. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 1 (2019). doi:10.1038/s41591-019-0462-y

25. Dou, Y. *et al.* Proteogenomic Characterization of Endometrial Carcinoma. *Cell* (2020). doi:10.1016/j.cell.2020.01.026
26. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Thirty-First AAAI Conf. Artif. Intell.* (2017).
27. Shia, J., Black, D., Hummer, A. J., Boyd, J. & Soslow, R. A. Routinely assessed morphological features correlate with microsatellite instability status in endometrial cancer. *Hum. Pathol.* **39**, 116–125 (2008).
28. Yamashita, H. *et al.* Microsatellite instability is a biomarker for immune checkpoint inhibitors in endometrial cancer. *Oncotarget* **9**, 5652–5664 (2018).
29. Vahadane, A. *et al.* Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Trans. Med. Imaging* **35**, 1962–1971 (2016).

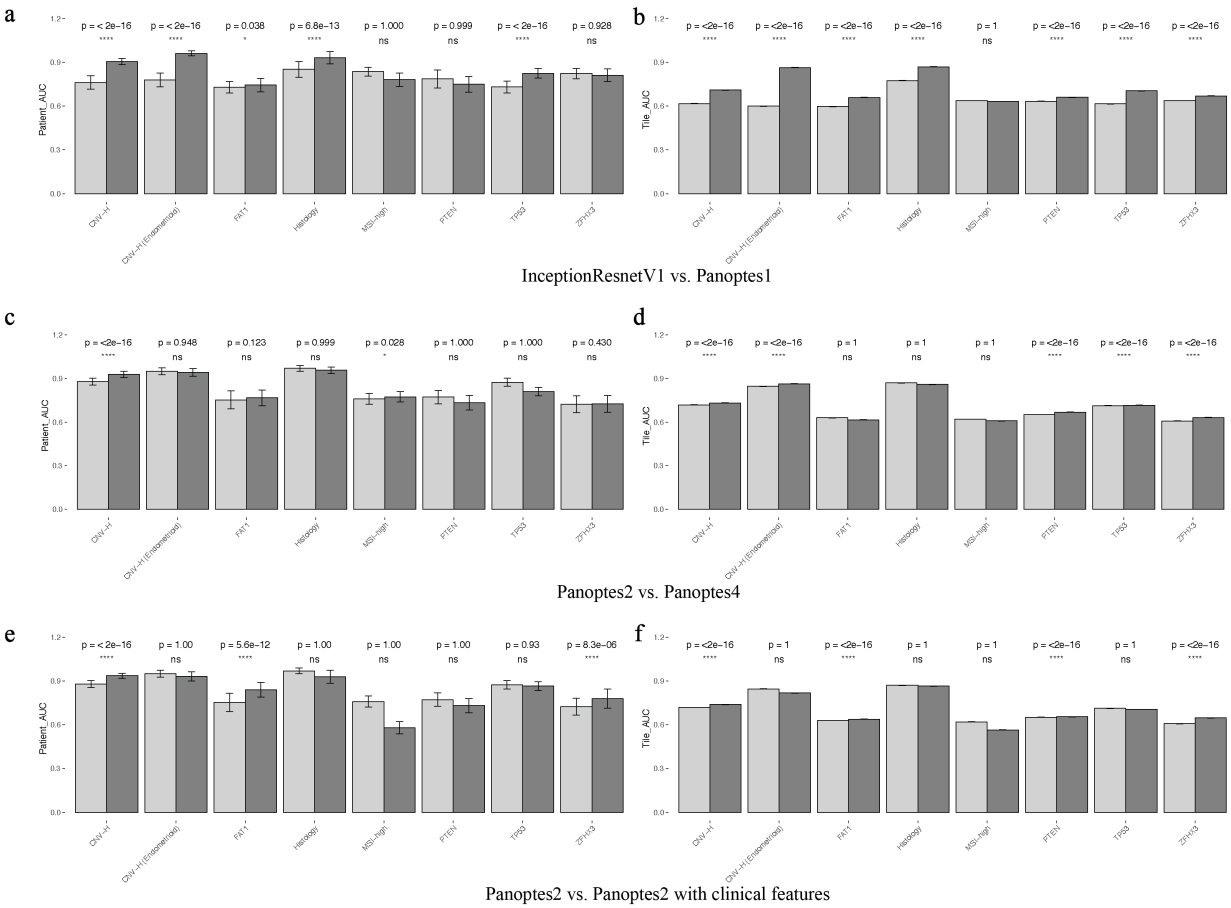
Supplementary figures



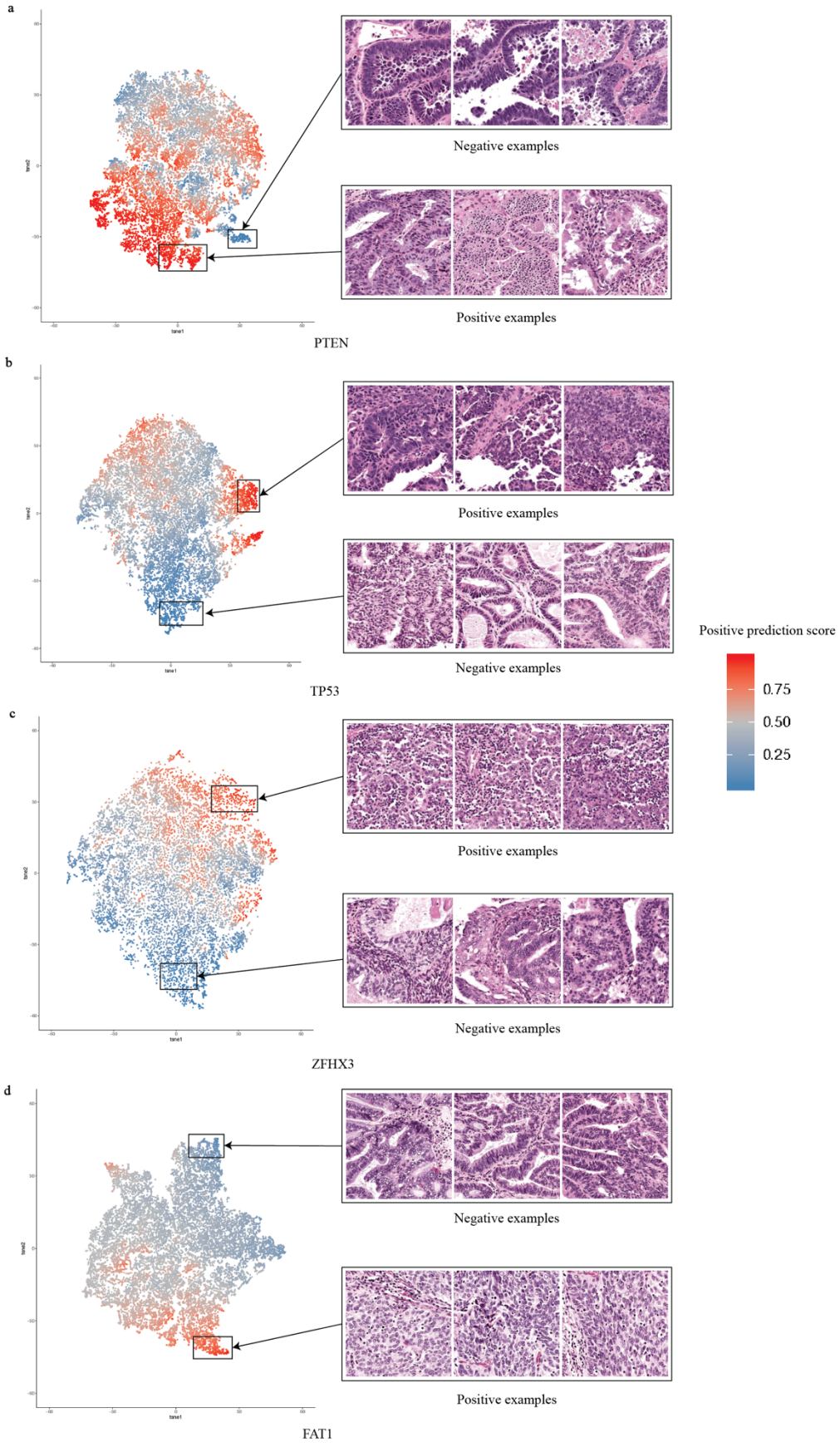
Supplementary Fig.1 | Data summary. a, Number of patients and composition of true labels in each task. **b**, Number of slides per patient in the cohort. **c**, Coefficient of colligation between subtypes and mutations. **d**, dimensions of slides in pixel (black: height; grey: width). **e, f, g**, Number of tiles per slide at 10X (**e**), 5X (**f**), and 2.5X (**g**) equivalent resolution.



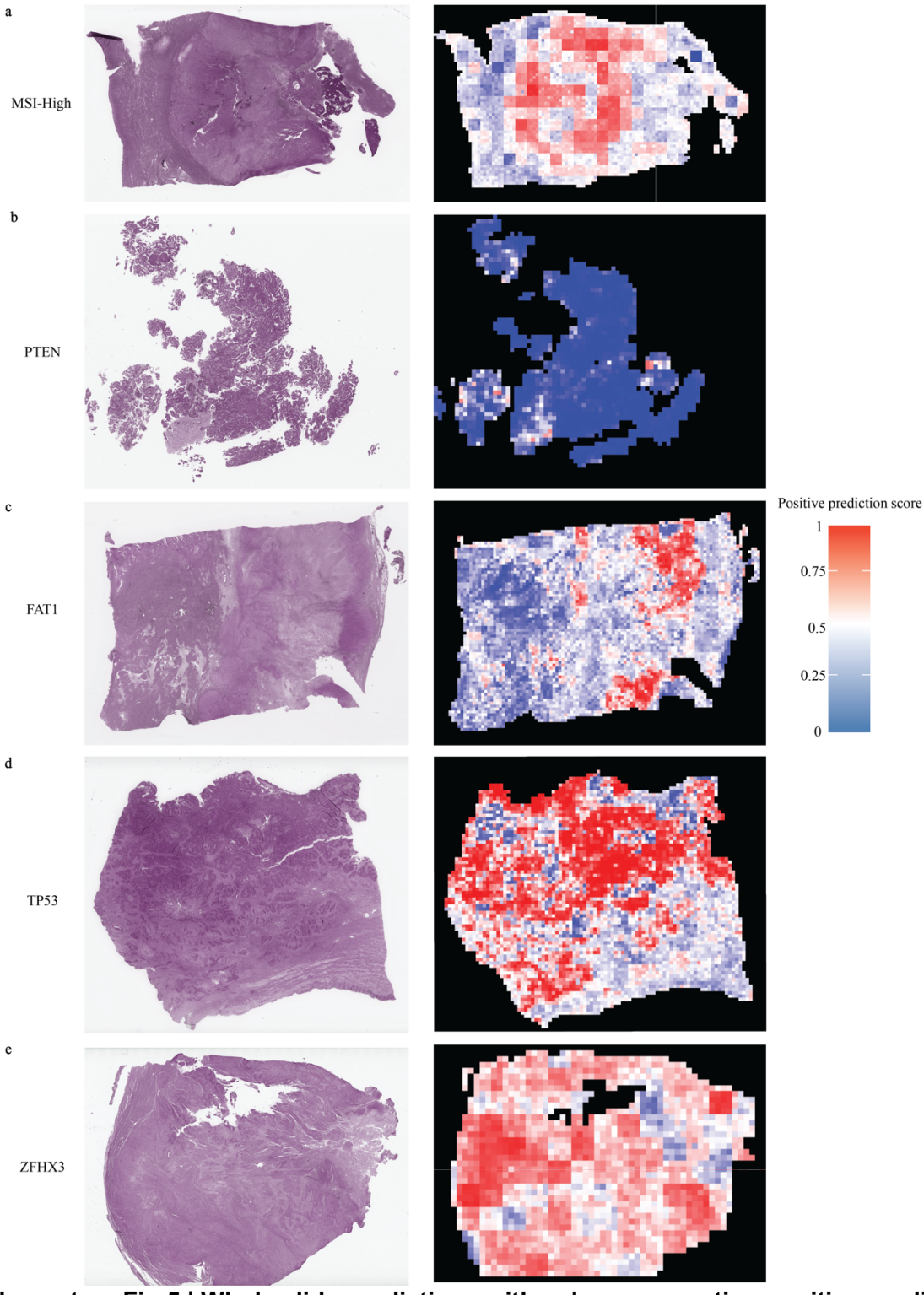
Supplementary Fig.2 | Two data split methods (a) and AUROCs achieved for Panoptes and baseline models on each prediction tasks using mixed random data split (b, c) and cohort independent data split (d, e) at per-patient and per-tile level. P represents Panoptes, PC represents Panoptes with clinical features, I represents Inception, and IR represents InceptionResnet.



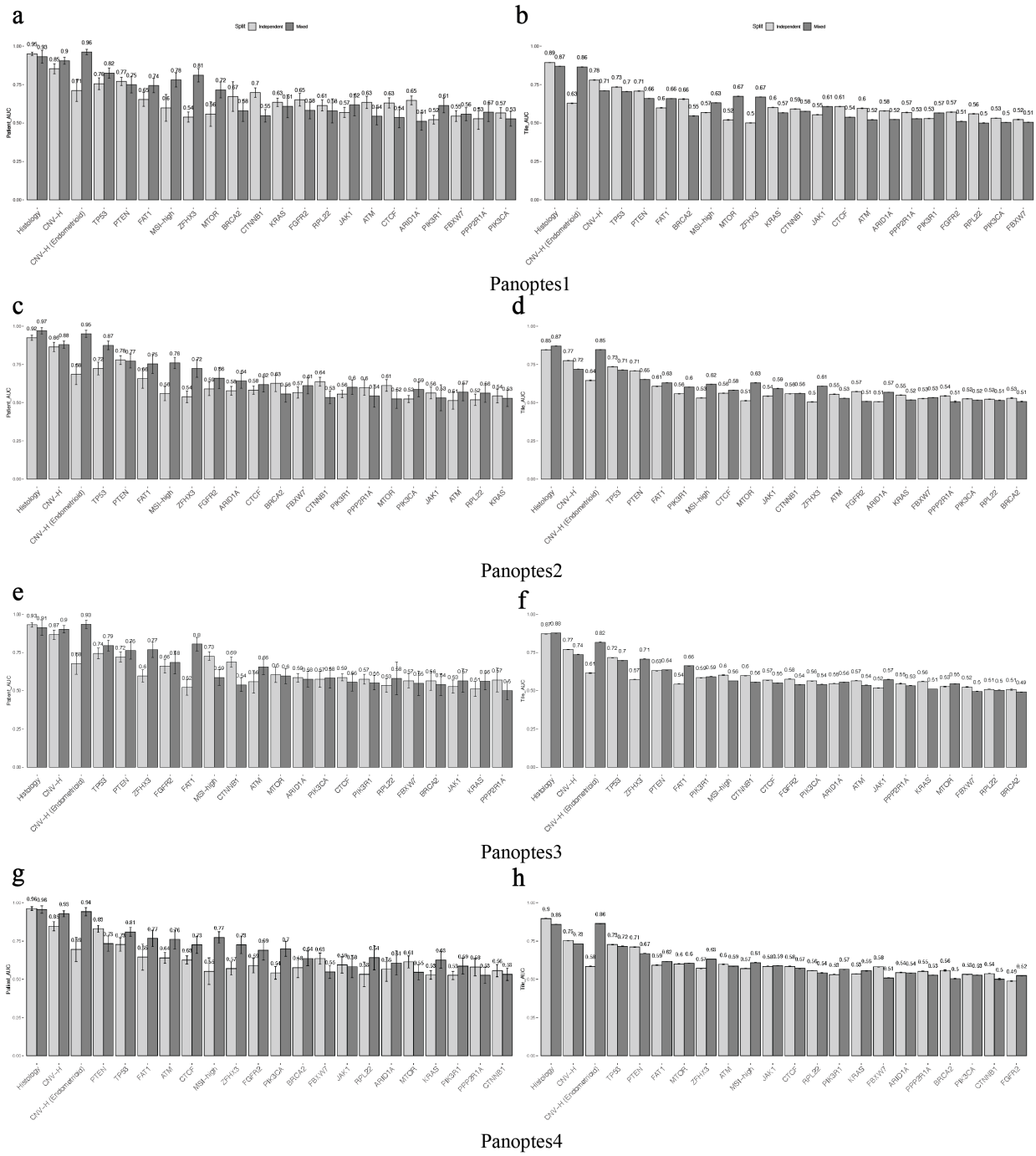
Supplementary Fig.3 | Comparisons of AUC between architectures on the top eight prediction tasks. a, b, 1-tail t-test of per-patient (**a**) and per-tile (**b**) AUROC between InceptionResnetV1 (light) and Panoptes1 (dark) of the top eight tasks. **c, d**, 1-tail t-test of per-patient (**c**) and per-tile (**d**) AUROC of Panoptes2 (light) and Panoptes4 (dark) of top eight tasks. **e, f**, Bootstrapped per-patient (**e**) and per-tile (**f**) 1-tail t-test of AUROC of Panoptes2 (light) and Panoptes2 with clinical features (dark) of top eight tasks.



Supplementary Fig.4 | Extraction and visualization of features learned by the models with tSNE. Each point represents a tile and is colored according to its corresponding positive prediction score. **a**, *PTEN* from a Panoptes2 model. **b**, *TP53* from a Panoptes2 model. **c**, *ZFHX3* from a Panoptes1 model. **d**, *FAT1* from a Panoptes2 with clinical features model.



Supplementary Fig.5 | Whole slide predictions with color representing positive prediction scores. a, Slide from a MSI-High (positive) patient using a Panoptes1 model. **b**, Slide from a *PTEN* wild-type (negative) patient using a Panoptes2 model. **c**, Slide from a *FAT1* mutated (positive) patient using a Panoptes3 model. **d**, Slide from a *TP53* mutated (positive) patient using a Panoptes2 model. **e**, Slide from a *ZFHX3* mutated (positive) patient using a Panoptes1 model.



Supplementary Fig. 6 | Side-by-side comparisons of AUROC between the Panoptes models in mixed random split trials and independent cohort split trials. Per-patient and per-tile level AUROC of Panoptes1 (**a, b**), Panoptes2 (**c, d**), Panoptes3 (**e, f**), and Panoptes4 (**g, h**) models in each task with mixed random data split (dark grey) and the cohort independent data split (light grey).